THE UNIVERSITY
of EDINBURGH

Perceptual Categorisation, Bayesian Inference and Psychological Similarity

Nina Poth

PhD Philosophy
The University of Edinburgh
2020

Declaration

I declare that this thesis is my own work, and has not been submitted for any other professional degree or qualification:

Nina Poth Bochum, 30th March 2020

Acknowledgements

Doing a PhD was not easy and the support that I have received from my supervisors and colleagues in the School of Philosophy, Psychology and Language Sciences at the University of Edinburgh and my family and friends made all the difference to what I could learn and accomplish during the three and a half years that this journey took.

First and foremost, I would like to thank my supervisors—Mark Sprevak, Alistair Isaac and Peter Brössel—for their guidance and helpful feedback on earlier drafts of this thesis. Thanks to Peter, I felt inspired to pursue a PhD in the first place. Thanks to Alistair and Mark, I was encouraged to pursue a career in philosophy. Thank you all for listening to my own thoughts, adding clarity and structure to my research ideas, spotting weaknesses in my arguments and helping me to grow in my writing and thinking skills.

My work has benefited from incredibly helpful comments by my peers and friends in the Departments of Philosophy and Linguistics at the University of Edinburgh: Svenja Wagner, Fausto Carcassi, George Deane, Kate Nave, Julian Hauser, Tamar Johnson, Amit Zac, Laura Jimenez, Matt Sims and Nicholas Rebol. I would also like to thank the various organisers and attendees of the *Postgraduates Work in Progress Seminar* and the *Mind and Cognition Group* at the University of Edinburgh. It was a lot of fun to discuss and learn about each other's work.

My special thanks belong to my friends and family. Yuanyuan was the best office mate I could imagine. Tamar, Amit, Julian and Livi have built a recharging home with me. I thank Franklin for taking me out on beautiful hikes in the Highlands. I'm grateful to Matt, Laura, Insa, Becky and Maddy for their open ears and kind hearts. I thank my family and friends in Germany for being 'rocks' in my life and my partner, Nicholas Rebol, for his endless love and care (and not to forget, the best pancakes I ever had).

Last but not least, I was able to do my PhD in Edinburgh due to the generous funding from the School of Philosophy, Psychology, and Language Sciences, the Royal Institute as well as the Aristotelian Society. The University of Edinburgh has offered me a great research environment, where I found many chances to openly talk about my research ideas. I am very grateful for this support and hope that some of the discussions in this thesis or those developing from it will redistribute this value in other ways.

Abstract

At the heart of this thesis is the following question: why do we categorise two objects (e.g., an apple and a banana) as instances of the same concept (e.g., the concept FRUIT) despite their perceptual differences? This is the problem of perceptual categorisation. One way of dealing with this problem is to appeal to the notion of psychological similarity: the apple and the banana belong to the same concept because they look similar. However, there is no scientific agreement on what entity or mechanism the notion of psychological similarity refers to and how this notion explains our ability to categorise both objects as fruit. A promising alternative approach to the problem is Bayesian modelling, whereby perceptual categorisation is analysed as a generalisation and concept-learning task: when categorising the apple and the banana as FRUIT, we compute the conditional probability that the banana is an instance of the concept FRUIT, given the background knowledge that the apple is an instance of this concept.

This thesis argues for a combination of a Bayesian and a similarity-based approach to perceptual categorisation. I argue that a Bayesian model of concept learning by Tenenbaum and Griffiths (2001) can help us to comprehend a variety of behaviours associated with perceptual categorisation. These were difficult to understand in light of two previous competing theories of psychological similarity—Shepard's (1987) geometric and Tversky's (1977) feature-matching theories. One of the behaviours that the Bayesian model can help us comprehend is the tendency to, for example, seek out mushrooms that look similar to edible ones and avoid those that look different from edible ones. The Bayesian model can help us understand why this tendency becomes stronger or weaker depending on how similar or different the mushrooms are. Another of these behaviours is a 'directionality effect': we are sometimes more likely to judge Tel Aviv to be similar to New York than vice versa. I argue that the Bayesian approach predicts, systematises and summarises the data on both types of behaviours, whereby it becomes a useful tool to understand perceptual categorisation as a unified phenomenon.

The second argument is that the advocated Bayesian approach implicitly relies on a theory of psychological similarity when characterising the hypotheses in the Bayesian inference of perceptual categories. The role of such a similarity-based theory is to explain how a concept such as FRUIT should be represented in a Bayesian model and how this concept's representational content is active in producing the subjective probabilities that are associated with hypotheses in a Bayesian inference task.

Lay summary

My thesis is a philosophical analysis of categorisation. For example, when we see an apple and a banana, we categorise them both as fruit. Why? I look at two types of answers to this question. One answer is similarity-based: we categorise the apple and the banana both as fruit because they taste similar (e.g., both are sweet). The other answer refers to probabilities: we categorise the apple and the banana as fruit because when we find apples, it is probable that we will find bananas as well (e.g., when we are in the supermarket).

It is not clear how these two kinds of answers can be combined. In my thesis, I propose one way in which we can combine a similarity-based answer with a probabilistic answer to the question of why we tend to categorise our perceptual experiences in this way. In Part I, I contrast and compare two fundamental models of similarity in computational-cognitive psychology—Shepard's (1987) geometric model of similarity and Tversky's (1977) feature-matching model of similarity. I make explicit that there is a conflict between the theoretical assumptions of these models, which leads them to apparently incompatible empirical predictions about similarity and categorisation. For example, one of these predictions is that our tendency to categorise two objects as the same decreases exponentially with a decrease in their similarity, a finding which is commonly referred to as the 'Universal Law of Generalisation'. The other prediction is that similarity and categorisation depend on the context, and in particular, on the direction of a comparison. For example, people typically judge Tel Aviv to be more similar to New York than vice versa. Neither of the theories of similarity can explain both of these findings at the same time.

In Part II, I evaluate a Bayesian model of concept learning by Tenenbaum and Griffiths (2001). I defend three philosophical criteria of unification and argue that this model meets these criteria. Thereby, it unifies these predictions and resolves the apparent incompatibility between them. On this basis, I propose that categorisation can be studied as a unified phenomenon.

Contents

Preliminaries

List of abbreviations

PC: Perceptual categorisation

PS: Psychological similarity

T&G: Tenenbaum & Griffiths

C&H: Colombo & Hartmann

G&K: Guttman & Kalish

ULG: Universal Law of Generalisation

MDS: Multi-Dimensional Scaling

Notational conventions

Throughout the monograph, and if not otherwise specified, I use small capitals to refer to concepts. I use single quotation marks to indicate oral, written or gestured expressions. I use italics to express emphasis. For instance, if I want to stress the difference between concepts and categorisation I write '*concepts* are mental representations whereas *categorisation* is the performance of grouping perceptual experiences into kinds that can form such mental representations.' Except for block quotes, I use double quotation marks to quote.

Preliminary remarks and terminological clarifications

Perceptual categorisation versus perception

The difference between perceptual categorisation (PC) and perception can be analysed as a distinction between two problems. PC is the problem of learning a cognitive category, where this representation also contains information about possible future members of the category. To approach the problem of PC, an agent must compare perceptual experiences associated with at least two perceptual stimuli. For example, categorising an apple as edible involves deciding that one can possibly eat it or that it *will* be edible, based on *previous* experiences with

apples that were edible as well. Thus, PC requires the imagination of possible states of objects that belong to a category (what edible objects are like) on the basis of comparisons across the perception of individual instances of a category (edible apples). This description of a perceptual category comes close to Carnap's (1988) notion of the intension of a concept.

This 'imagistic' aspect of PC can be detached from the immediate perceptual experience that comes with object perception, which is the problem of finding a stable representation from a mess of varying sensory inputs to the perceiving system (e.g., the visual system). For example, perceiving an apple requires, among other things, representing the constancy of its colour and distinguishing its figure from its surrounding background (Wertheimer, 1923/2013). In this sense, the ability to perceive is bound to the immediate perceptual experience of a perceptual stimulus (e.g., the apple) but PC is not.

There are two sorts of relations between PC and perception. On the one hand, PC directly and necessarily relies on perceptual abilities but perception does not directly rely on the ability to perceptually categorise. For example, it is impossible to categorise a mushroom as edible without ever having tasted a mushroom in the past whose edibility can be taken as a standard of comparison for future instances of edible mushrooms. On the other hand, it is possible to perceive a mushroom without categorising it as edible.

On the other hand, PC can facilitate perception in the sense that it makes perception more discriminative. A pattern of air pressure might be sensed with receptors in the cochlea (and hence processed in the nervous system), while not being (consciously or unconsciously) identified as a sound. The phenomenon of perceiving the pattern of air pressure as a sound such as a voice or a tone has also been described as *categorical perception* (cf. Harnard, 1987). Categorical perception increases the discriminability of different stimuli. Practice in categorising musical pieces and tones increases the ability to perceive a tone as a 'C' as opposed to an 'E'. Such classifications commonly require the stimuli to achieve a certain threshold (e.g., in loudness or pitch) to be categorically perceived.

Like perception, PC is not limited to human cognition. An example for this comes from a famous study by Guttman and Kalish (1956a) (G&K), who have shown that pigeons can be trained to generalise a pecking response from a key lit by some wavelength x to a key lit by a similar wavelength y. A door would subsequently open, presenting a container of seeds to the pigeons. This finding has typically been interpreted to mean that these pigeons categorise x and y as the same perceptual inputs (see also chapter 3.2).

One way in which the achievement of PC is typically explained is on the basis of mental representations. For example, G&K's explanation of pigeon's behaviour is that the pigeons compare the two differently lit keys mentally with each other and build a stable representation that extracts the information associated with each of the keys for reaching the food before the door opens. This style of explaining behaviour based on representations follows the sandwich model (see *perception* entry in glossary). When generalising, the pigeons in G&K's study

(a) perceive the new wavelength, (b) compare it with the old wavelength (e.g., based on an internal representation of their similarities) and (c) output a pecking response. PC involves all stages, (a), (b) and (c) but perception is the process limited to stage (a). Although this picture is very simplified (as it is clear that the particular processes underlying PC must be more complex), it has led to remarkable computational models today, which have helped to systematically investigate PC. Some of these models will be discussed in this thesis and it will be argued that these models help us to better understand PC.

Staying neutral about perceptual categories

This thesis is not about the question whether perceptual categories are metaphysically real (i.e., beyond the senses). There might not be a real kind in the world that corresponds to the things that we call 'apple' or 'mushroom'. This thesis takes also no position towards whether perceptual categories are real (i.e., in the world).

I view perceptual categories as cognitive representations. What a perceptual category represents depends at least in part on its purpose for adaptive behaviour. For example, an organism's representation of the category of red things is what it is because of the significance of this category for the organism's survival (e.g., it typically correlates with dangerous events). Behaviour can be adaptive even if it happens on the basis of a hallucinated representation that does not correspond to anything real in the world. For example, an agent might be very successful at avoiding poisonous mushrooms and seeking edible mushrooms despite her hallucinatory representation of the mushrooms as apples. A cognitive agent's representation of the class of all edible mushrooms in the world need not correspond to properties of the set of all edible mushrooms in the world. It just needs to represent for the agent relevant relations between things in the world that cognitive agents would treat as mushrooms under a given set of environmental conditions.

Perceptual miscategorisation

Representations of perceptual categories must allow for misrepresentation, and whether or not they do depends on whether they lead to successful or unsuccessful behaviour. PC fulfils the function of guiding action qua representing relevant information in the world. For example, categorising two lit keys as food-relevant evokes a pecking response to both of them qua representing their similarities in their light intensities. If the pigeon categorises the second key as too different from the first key and avoids a pecking response, no door will open and no food will be served. This is a case of miscategorisation as it prevents the pigeon to obtain food.

Here I assume that the conditions for miscategorisation are observer-relative. Why some pieces of information about the world are relevant for comparison is

decided in retrospect by an observer's (e.g., a scientists) evaluation of whether an action was adaptive to a given environment or not. To identify miscategorisation, an agent's behaviour has to be evaluated in retrospect with respect to its adaptive value.

Perceptual categorisation and adaptive value

In this thesis, I assume that all cases of perceptual categorisation potentially involve adaptive value. That is, I assume that the ability to categorise groups of objects according to internal representations or concepts potentially has consequences (even if only over the long run) for the thriving and survival of a perceptual categoriser in its environment. For example, there is a great advantage in being able to infer which mushroom exemplars are edible and which ones are poisonous. Inferring the correct one of these categories is crucial for the agent to decide what behavioural response (e.g., eating or avoiding to eat) should be given to the objects. The behavioural response, in turn, will (at least partly) determine how likely the agent will be to thrive and survive—eating edible mushrooms feeds the agent with important nutrients, while eating poisonous mushrooms may disturb or even kill the agent. Thus, the adaptive value associated with the ability to perceptually categorise objects (e.g., edible mushrooms) as belonging to a category or concept (e.g., EDIBLE) is that it increases one's chances of having a longer or better live. Some possible perceptual categorisations may be of disvalue, for instance, when they decrease one's chances to survive or thrive (e.g., the identification of poisonous mushrooms as edible).

Why should we consider a criterion of adaptive value for the norms of PC behaviour? The role of adaptive value in this thesis is particularly to help explain why PC is sensitive with respect to different contexts and environmental niches. For example, developing a category system that involves the concepts EDIBLE MUSHROOM and POISONOUS MUSHROOM is context sensitive in the sense that it facilitates behaviour adapted to 'natural' environments in which the sole goal of behaviour is survival. Learning to eat edible and avoid poisonous mushrooms serves this goal as it will increase an agent's chances to thrive and survive. In contrast, a more complex category system that also includes subordinate concepts such as FLY AGRARIC MUSHROOM, PORTOBELLO MUSHROOM, CHANTERELLE MUSHROOM etc. seems to facilitate behaviour that is adaptive to a different environment (e.g., a cooking competition or a biology classroom). In this case, the goal seems to be different (e.g., to select the most suitable mushroom for a delicious meal or to learn about biological distinctions among mushrooms). In the latter case, behaviour that is only selective between edible and poisonous mushrooms would carry less adaptive value with regards to the latter environments—it simply is not conducive for the relevant goal. Thus, the role of adaptive value for PC is to determine the goal or context with regards to which PC behaviour is successful (or not).

A reason to question the criterion of adaptive value is that some cases of categorisation seem to be, at least intuitively, unrelated to adaptive value (e.g., a classification of Whales as mammals). However, it is, in the first place, unclear whether the classification of Whales as a mammals is a genuine case of perceptual categorisation. For centuries, Whales and Dolphins had actually been classified as fish. The newer insight that they should be classified as mammals is based on a closer examination of their essential features (e.g., their hot-bloodedness, their ability to communicate) as revealed by the natural sciences. In contrast, the earlier categorisation of Whales as fish seems to be closely related to their appearance (e.g., their fish-like shape and the observation that they live in water). It is unclear that the novel categorisation is a case of perceptual categorisation, whereby it is not ruled out that the earlier classification may have had some adaptive value. There seems to be a difference between perceptual (e.g., 'Whale', based on its perceptual properties) and non-perceptual (e.g., 'Whale', based on a scientific taxonomy) cases of categorisation. While category systems in scientific taxonomies seem to follow scientific norms (e.g., norms that concern how well instances of these categories can be identified on the basis of their scientifically observable features), I assume that perceptual categorisation follows the norms of adaptive behaviour; Perceptual categorisation is of potential adaptive value in all cases. Whether the more general cases of categorisation, such as in scientific taxonomies, have the potential to be of adaptive value as well cannot be concluded by the work in this thesis.

Psychological similarity versus similarity of natural kinds

Psychological similarity contrasts with the similarity of natural kinds. At least two differences between the two stand out. Firstly, the similarity associated with natural kinds concerns groupings of objects in nature. For example, members of the kind electron can be grouped into those that are positively charged and those that are negatively charged. In contrast, psychological similarity concerns groupings of objects and their associated properties that are accessible to the perceptual organs of organisms. Electrons are not perceivable, neither are kinds of colours, such as ultraviolet. Secondly, the two domains differ with respect to how they ground similarity. One approach to grounding natural-kind similarity is to derive it from essences: members of a kind are similar because they have the same essence (e.g., being of the substance H_2O). In contrast, psychological similarity is grounded in perception, cognitive function and survival. For example, stones with different minerals such as jadeite and nephrite are psychologically similar because they are perceived as the same kind of stone—Jade. Thirdly, the similarity of natural kinds is objective. In contrast, psychological similarity depends on the subjective aspects of how a person sees the world. This last point locates questions about natural-kind similarity in the area of fundamental questions about science: the similarity of natural kinds is a basis for scientists to draw distinctions and classify their scientific discoveries in a way that reflects information about nature or reality. Natural-kind similarity is a basis for scientific

inductions based on scientifically objective observations, whereas psychological similarity is in the area of questions of psychology; in particular, questions about the functioning of psychological processes. Psychological similarity is a basis of induction by humans and other animals based on subjective experiences.

Similarity versus probability

Similarity in the content of this thesis is psychological similarity. Two objects are similar if they are perceived to be similar and two political regimes are similar if they are conceived of as similar. Thus, psychological similarity is a property bound to perceiving/conceiving subjects (or groups thereof). Two people might perceive two objects to be differently similar, and this will depend on their different conceptual or perceptual representations of the object. Thus, similarity is a mind-dependent property.

Despite their differences in detail both Shepard's and Tversky's theories explain PC as a result of cognitive processes that calculate similarities. For instance, apple and a banana are both in the category *fruit* because they are similar enough, as compared to other things that fall into different categories.

T&G's theory, on the other hand, explains PC as a cognitive process that calculates probabilities instead of similarities. For instance, apple and a banana belong to the same category, *fruit*, because it is more likely to observe them together given that they are members of that category. The psychological similarity- and the probability-based explanations use different explanans. Either the probability to which two or more items are placed into the same category depends on their psychological similarity or it depends on some probabilistic process itself. According to the latter explanation, PC involves probabilistic inference. According to the former, similarity explains PC. The debate is about which of these explanations of PC is better, and according to which criteria this can be evaluated.

Motivated by the Shepard & Tversky debate about a coherent account of psychological similarity, T&G have argued that a probabilistic explanation is superior to a similarity-based explanation because (a) it is simpler and does not need an additional concept of psychological similarity, (b) it is more general because it can unify both Shepard's and Tversky's accounts, and (c) it can predict empirical effects that could previously only be explained by either Shepard's or Tversky's theories but not by both. In effect, T&G have suggested that psychological similarity is an unnecessary concept in the explanation of PC and should hence be abandoned.

In my PhD thesis, I refute T&G's case against psychological similarity and argue that there is yet no reason to replace an explanation via psychological similarity by an explanation via probabilistic inference. This argument is supported by my illustration that the probabilistic explanation itself relies on a theory of similarity. My alternative proposal is that a Bayesian (i.e., probabilistic) approach to PC can offer 'what-if' explanations (van Rooij, Wright, Kwisthout,

& Wareham, 2018) that can describe aspects of possible psychological processes (e.g., similarity-based processes) in particular individuals that seem to generate behaviour associated with PC. I argue that such an approach can interact with similarity-based explanations but it should not be taken to replace them.

1. Introduction: Modelling perceptual categorisation

1.1. The problem and the claim

Perceptual categorisation (PC) is the ability to generalise behaviour from old perceptual experiences to new perceptual experiences. For example, upon having eaten an umami portobello mushroom, and in light of the new experience of a bitter fly agraric mushroom, an agent will be likely to seek further instances of portobello mushrooms and avoid instances of fly agraric mushrooms. The problem is how to explain this ability to generalise.

In this thesis, I approach this problem from a computational-modelling perspective. This perspective contrasts with a neuroscientific approach to explaining PC. The goal of my thesis is to present a *unified* approach to PC. The key claim of my thesis is that a Bayesian explanation of generalisation, inspired by Tenenbaum & Griffiths' (2001) Bayesian model of concept learning, can deliver such a unified approach. I contrast my contribution against two previous attempts to explaining behaviour associated with PC in terms of a psychological model of similarity. These attempts are Shepard's (1987) model of geometric distance and Tversky's (1977) model of feature-matching.

These previous attempts had treated the behaviour that we commonly associate with PC as an effect of two separate kinds of cognitive processes—one process calculates geometric distances while the other matches features. The models associated with each of these assumed processes are experimentally well confirmed. Problematically, Shepard's and Tversky's theoretical assumptions about how psychological similarity explains the observed behaviour are inconsistent with each other, so that it is difficult to combine their explanations of how a unique psychological process (e.g., one of similarity) could generate the observed PC behaviour. At the same time, because of the solid empirical evidence for Shepard's and Tversky's models, it is difficult to decide which of their theories comes closest to a correct explanation of PC. Does the behaviour associated with PC at all express a unique internal psychological mechanism? In this thesis, I approach the conflict between these competing theories of psychological-similarity processes from a novel, Bayesian, perspective. The principal virtue of my approach is that it offers a computational explanation of PC as a unified phenomenon.

1.2. Positioning the approach

On some accounts in the psychological literature, it is common to view PC as an explanans when answering questions about related cognitive phenomena. An example is Bundesen's (1990) study of subjects' behavioural performance in visual identification tasks. The task is to identify digits on a display where the digits have different colours. When interpreting the data, Bundesen refers to subjects' cognitive ability to discriminate the digits based on the colour-category that the digits are assigned to. Correspondingly, it is this ability that influences the speed and accuracy of subjects' performances (Bundesen, 1990, pp. 523–524). The observation that red digits are commonly identified more quickly than digits of other colours is explained by the assumption that red digits are more relevant to the subject for solving the task than digits with other colours (e.g., because red is often associated with concepts such as DANGER). Thus, assuming that the subject engages in a PC task explains the subject's performance when identifying digits. There are other approaches that consider PC as an explanans when studying concept- and word-learning (e.g., Sloutsky, 2010; Smith & Samuelson, 2006).

In contrast to this literature, *this thesis takes PC to be the explanandum.* From this perspective, there are three relevant questions about PC:

1. Why should we do PC in the first place?
2. Why do we make the specific perceptual categorisations that we do?
3. What is the relationship between PC and similarity?

My motivation to consider PC as the explanandum is that there seem to be no clear answers to these questions in the psychological literature, revealing that PC is itself not well understood. In this thesis, I propose that a Bayesian approach offers partial answers to these three questions. In answering the first question, this thesis focuses on a computational approach to PC. This approach is inspired by Marr's (1982) computational level analysis of information-processing systems and Anderson's (1991; 1991a) rational analysis of adaptive organisms, which will be reviewed in chapter 2. My approach to PC can be positioned alongside these approaches as it also investigates the psychological process concerning possible algorithms and representations used to successfully perform PC. In answering the second question, I focus on the assumption in the literature that PC is guided by certain principles of rationality, based on which some perceptual categorisations appear to be better than others. My answer to the third question is that the Bayesian approach to PC that I advocate here implicitly relies on a theory of similarity, when this approach is used to explain the psychological processes and concepts that drive aspects of behaviour associated with PC.

Previous answers to the three questions have often built on two fundamental studies about categorisation and similarity. These are Shepard's (1987) *Universal Law of Generalisation* (henceforth 'ULG') and Tversky's (1977) feature-matching model of similarity. In these studies, categorisation is explained by similarity. For example, the explanation of why we categorise apples and bananas as instances

of the fruit category is that they are similar to each other. The problem with this sort of explanation is that the notion of similarity is itself poorly understood; it is not clear what 'similar' in such explanations means (Decock & Douven, 2011; Goodman, 1955). A popular alternative to similarity-based explanations of categorisation are probabilistic explanations. An early example are models of cue validity, which represent aspects of categorisation in terms of the conditional probability of a category given a cue (Reed, 1972; E. Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). The corresponding explanation of why we categorise apples and bananas as instances of the fruit category is that it is more probable that apples and bananas belong to the same category given that they often occur together with the same cues. Recently, probability-based explanations of categorisation have been taken up again, e.g., in a Bayesian model of concept learning by Tenenbaum and Griffiths (2001). It is still unclear how these two types of explanations of candidate psychological processes underlying PC behaviour can be related.

What sets my project apart from these previous approaches to similarity and categorisation is that it explicitly connects a Bayesian (i.e., one kind of probabilistic) and a similarity-based explanation of PC. Instead of choosing between the two conflicting approaches to psychological similarity, I suggest to take on a novel, unifying, perspective on PC. This third approach to PC unifies two different empirical phenomena associated with similarity and categorisation according to three philosophical criteria of unification. One phenomenon consists in Shepard's (1987) observation that the behaviour associated with PC shows an *exponential gradient*: the probability to generalise from one experience to another decreases exponentially with the perceived similarity between these experiences. The other phenomenon consists in Tversky's (1977) observation that judgements of similarity often show an *effect of directionality*. For example, Tel Aviv is typically judged to be more similar to New York than vice versa. Earlier, no philosophical argument for unification of these phenomena had been provided and it was not clear how these phenomena could be unified.

In this regard, this thesis does not refute either of the similarity-based explanations. With regards to questions about the psychological processes and representations, the competition between the two theories of psychological similarity remains. Which of these theories of psychological similarity is ultimately correct cannot be decided on the basis of my approach alone, which only grounds a third possible theory of these phenomena.

My proposed unified approach is useful beyond the current state of the art on the problem of modelling PC in two respects. Firstly, in clarifying the explicit connection between a Bayesian and a similarity-based explanation, the unified approach helps us to better understand the relationship between questions about psychological-similarity processes and the apparent guiding principles of behaviour associated with PC. Secondly, the unified approach clarifies Tenenbaum and Griffiths' (2001) earlier Bayesian approach as a computational theory, its possible limitations and its possibility to contribute to the Shepard-Tversky

debate about psychological-similarity processes. This helps us to comprehend the diversity of behavioural patterns associated with PC.

1.3. The overall argument

The general structure of my argument that a Bayesian approach unifies behaviours associated with PC (and that this is worth considering) is as follows.

1. There are two candidate approaches to explaining patterns of behaviour associated with perceptual categorisation. On the one hand, there is Shepard's (1987) geometric approach to similarity and his Universal Law of Generalisation (ULG). On the other hand, there is Tversky's (1977) feature-matching model of similarity.

2. These approaches can accommodate different patterns of behaviours. Shepard's approach can accommodate the exponential gradient, while Tversky's can accommodate the effect of directionality. However, they cannot agree on a unique psychological process that generates these types of behaviours. They conflict with regards to their theoretical assumptions about what similarity is.

3. An alternative approach to explaining these different patterns of behaviours is offered by Tenenbaum and Griffiths' (2001) Bayesian model of concept learning. Principally, the Bayesian model proposed by Tenenbaum and Griffiths is neutral about the conflict between Shepard's and Tversky's assumptions about psychological similarity processes.

4. The Bayesian approach satisfies three criteria of unification: (i) it is simple because it uses only a few formalisms to analyse PC, (ii) it has a broad scope because it predicts both, the exponential gradient and the effect of directionality, and (iii) it reveals that these different patterns of behaviours appear to be dependent on each other, while they had appeared to be independent before. In this sense, the Bayesian approach unifies the diversity of behaviours associated with PC.

This argument is embedded within the following structure of this thesis. Chapter 2 provides a working definition of PC and characterises the framework of reverse-engineering, which I assume as a background for the advocated Bayesian approach to explaining PC. In Chapters 3 and 4, I discuss Shepard's and Tversky's theories of psychological-similarity processes as candidate answers to questions 1-3 above. Chapter 5 states the conflicting assumptions of these theories and motivates my project of unifying their distinct empirical predictions.

Chapter 6 explains the Bayesian approach to PC that I advocate in this thesis on the basis of Tenenbaum and Griffiths' Bayesian model of concept learning. The chapter clarifies the implicit connection between T&G's model and Shepard's (1987) geometric theory of psychological similarity. Chapter 7 reflects on

the possible limitations of the Bayesian approach at the computational level of explanation. This explicates my claim that a Bayesian explanation of how the cognitive mechanism of PC could work implicitly relies on a similarity-based approach.

Chapter 8 argues for three criteria of unification that can be used to evaluate the Bayesian approach to PC: elegance, unbounded scope and informational relevance/probabilistic dependence. Chapter 9 shows that the advocated Bayesian approach to PC satisfies these criteria, and argues that PC is a unified phenomenon. I conclude with two questions for future research.

2. Background

This chapter presents the theoretical background of my approach. In section 2.1, I provide a working definition of the explanatory target—perceptual categorisation (PC). In section 2.2, I motivate the Bayesian approach to PC in the context of the reverse-engineering strategy in cognitive science.

2.1. The explanatory target

PC is the ability to recognise and distinguish objects on the basis of their perceptual attributes. Following Palmeri,

> [p]erceptual categorization is a fundamental aspect of human cognition. Any time we decide that some visually presented object is a dog rather than a cat, a bottle rather than a jar, or a tree rather than a shrub, we are making a categorization decision based on the perceptual attributes of that object. (Palmeri, 2001, p. 193)

Accordingly, PC is the cognitive ability to decide for any perceptual experience what category the experience belongs to. PC serves a variety of other cognitive processes, such as decision-making for the control of behaviour, recognition and learning. PC allows organisms to use stored information about past perceptual experiences in an appropriate way. For example, recognising a red traffic light as dangerous makes us more likely to avoid car crashes. Likewise, being able to distinguish a portobello mushroom from a fly agraric mushroom can help us to learn that one is edible but the other is not.

Five aspects stick out that together provide a working definition of PC: it is cognitive, possibly unconscious, hierarchical, normative and wide-scoped. I discuss each of these aspects in turn.

Cognitive

PC is a cognitive process and PC behaviour may be interpreted in terms of stimulus-response behaviour. Such behaviour may, for example, express aspects of

associative learning (e.g., classical conditioning[1]). Although some PC behaviour may express cognitive processes that build upon a background of coordinated automatic (unconditioned) response behaviour, PC behaviour is not limited to giving an unconditioned automatic response. Correspondingly, PC behaviour does not consist of inflexible action patterns (e.g., reflexes). PC behaviour is flexible and can be modified to suit the cognitive agent's needs.

An example illustrates this basic point. Mushrooms are known for being often confused as edible when they are in fact poisonous. When a dog eats one mushroom and avoids another, this is an instance of PC. The dog decides on the basis of the smell of one mushroom whether to sniff around it further and eat it or whether to avoid it. If the mushroom carries a very sweet smell, the dog will not eat it. If it smells fresh, the dog may eat it. This process is not a mere reflex-like response to the perceived smell because there are many different ways in which the perception of the mushroom could have been interpreted by the dog. More generally, the choice of eating or avoiding the mushroom is a choice among multiple alternative interpretations of a given situation.

To abstract from the example, PC involves thinking about relations between perceptual experiences and deciding whether they should be treated as belonging to the same or different perceptual categories. Intuitively, similarity plays a role to make this decision. In the mushroom case, the dog must decide whether the smell of the mushroom is like the smell of a previously experienced mushroom that was edible. To make a choice, some criterion of sameness between these experiences is necessary. If the mushrooms smell similar enough, they should be treated as the same—eat one and feel good, eat the other and you will feel good. If the mushrooms' smells are too dissimilar to each other, then the mushrooms should be treated differently—eat one and feel good, be cautious about the other because you might not feel good after eating it. Thus, the decision as to how perceptual inputs (e.g., two mushrooms) shall be treated (e.g., eaten) is not arbitrary, it is made on the basis of a criterion of similarity or difference (e.g., similarity in smell).

As we will see, a big question in research on PC is: what does 'similar enough' or 'too dissimilar' mean? Chapters 3 and 4 take this question to a slightly more general form: 'what is psychological similarity?'

[1] In classical conditioning, an unconditioned stimulus is repeatedly paired with a conditioned stimulus. For example, in Pavlov's original experiments on dog salivation, the unconditioned stimulus is a piece of meat, the conditioned stimulus is the sound of a bell and the salivation response is a conditioned response. The goal of classical conditioning is to make the subject (e.g., a dog) build an association between the unconditioned stimulus and the conditioned stimulus so that, whenever only the unconditioned stimulus is presented, the subject elicits the conditioned response. For example, the dog has learned an association between the bell ring and the meat when, upon hearing only the ring of the bell (i.e., without the actual presence of the meat), the dog salivates, while, prior to the conditioning, salivation was naturally elicited only upon the presence of the meat (as cited by LeDoux (1998, 142)).

Possibly unconscious

Perceptual categorisation is not confined to conscious processes, it can also happen unconsciously. For example, PC can be conscious, such as when subjects verbally report on the categorisation, or unconscious, where, despite behavioural measures indicating PC has taken place, subjects are neither consciously aware this is the case nor able to report on it. This claim is motivated by the idea that the ability to categorise objects, events or situations that are associated with our perceptual experiences is vital for decisions that we make on a daily basis. These decisions are often quick and need not involve conscious reflections or deliberate reasoning.

It is possible for a subject, S, to process information cognitively but unconsciously. A non-trivial view of consciousness assumes that to perceptually categorise a set of stimuli, S need neither be consciously aware of the presence of the stimuli nor be consciously aware of the possibilities for acting upon them/the fact that S is categorising them. S can be said to perceptually categorise objects into classes if S behaves distinctively in behavioural patterns towards the objects, without being able to verbally report that S is doing so (cf. Kouider & Faivre, 2017). This notion of unconscious PC does not preclude the possibility that subjects perceptually categorise also when they are able to consciously report on their discriminatory behaviour.

When investigating PC in terms of reliable behaviour patterns towards certain objects, we commonly assume that animals that are not able to report on the contents of their experience and lack linguistic competence are also able to perform PC. Thus, although research in categorisation has largely focused on the question of what aspects of our perception of the world determine the use of linguistic categories, questions about PC apply to non-linguistic animals as well. For example, one can ask: why does a mouse eat all the raisins but not the poisonous berries? In this sense, the ability to make differences in categorisations seems to be species-independent. This awareness has been raised by the psychologists Rosch and Mervis (1975) on empirical grounds and by Wittgenstein (2006, sec. 65-78), on empirical and theoretical grounds.

Hierarchical

Explaining PC typically requires reference to perceptual categories. PC is a cognitive process by which a set of distinct observations is grouped into one single abstract representation—what I call a cognitive category, in the style of E. H. Rosch (1973), or a concept, in the style of Fodor (1987). (Both terms have an entry in the Glossary.) Perceptual categories are hierarchically nested. For example, the observations of a terrier, a shepherd and a pug can be grouped into the category *dog*. At the same time, these observations are all instances of the categories *pet* and *mammal*. This illustrates that the perceptual experiences of these instances can be categorised at various levels of a taxonomic hierarchy (cf.

Gelman & Markman, 1987). The hierarchical structure of perceptual categories presents a problem of underdetermination. Given a set of instances of a category, there is no unique solution to the categorisation task.

One approach to solve this problem is to explain PC in light of a notion of PS. For instance, Rosch and Mervis have famously argued on an empirical basis that cognitive categories are structured in terms of the relative degree of PS among their members. The resulting representation in a categoriser's mind takes the form of a prototype.[2] According to Rosch and Mervis' prototype theory, an S should categorise an object as a bird if S's mental representation of the object is similar to its representation of other possible members of the *bird*-category. One way to interpret this constraint is to say that x is similar to the other members only if x shares enough properties with them (Margolis & Laurence, 1999, p. 29).

Taken together, the examples from this literature illustrate that the structure of cognitive categories or concepts is inherently nested and no clear boundaries may be defined between them. A penguin may be categorised as a penguin, as a bird, or as an animal. This is a problem for similarity-based accounts of PC. Although PS may be a guide to the structure of these categories, it does not explain why subjects sometimes categorise penguins as penguins and sometimes as birds or animals. An additional approach to finding out why subjects sometimes prefer some of these categories over others is needed. Resolving this indeterminacy is a motivation for placing a normative constraint, that is, a specification of why one or other category should be preferred given certain conditions. These conditions are best understood in light of the fact that PC is a solution to a functional problem that is performed relative to the subject's goal.

Normative

This problem can be characterised as the problem of treating stimuli appropriately. Take x to be a known perceptual stimulus and y to be a novel perceptual stimulus. For example, x is a portobello mushroom and y is a fly agraric mushroom. S knows that x is edible. The problem is: given our treatment of x, should S treat y in the same way and expect it to be edible? The task is to build a cognitive category that says whether, and if so then to which extent, a behavioural response that was given to an old perceptual input should be generalised to a new perceptual input. Upon eating x, should S eat y as well?

A generic solution to the task includes roughly two steps. (1) Take a candidate perceptual category, for instance *edible*. (2) Have a criterion to compare x and y. (3) Use the criterion to decide how to treat y. Together, (1), (2) and (3)

[2] A prototype is an item that is maximally similar to any random member of the category. For example, a robin is a prototype for a bird because robins are maximally similar to all other sorts of birds. They typically have many features in common with other birds (e.g., they fly, they have two legs, they are of a small size, etc.). On the contrary, a penguin is not a prototypical bird because, although it has two legs, it is neither of a decent size nor can it fly. Hence, the penguin does not share many attributes with the average of all other birds.

are abstract components of a function that transforms a number of perceptual inputs (mushrooms) into a category representation (e.g., *edible*) and attaches a behavioural disposition or response to it (e.g., eating).

Who determines what the problem is? One way in which the problem of PC can be identified is from a teleofunctional perspective, which can be observer-independent (cf. Millikan, 1989) or observer-relative (cf. Dretske, 1988). The observer-independent perspective assumes that the problem is set by the etiological history of the system that consumes the behaviour. In the case of PC, the consumer of S's categorisation behaviour, e.g., the eating, is S's body. From this perspective, the purpose of a PC mechanism is to cover the adaptation or survival of S as a whole organism. On this view, PC is relative to organismic needs and what these needs are is independent of an external observer (e.g., a scientist). Likewise, the normative constraint on PC is its adaptive value, which is not up to us but to processes in evolutionary history. From this perspective, when a dog decides whether to eat or avoid a mushroom, it should act upon only those perceptual categories (e.g., those containing only edible mushrooms) that increase the adaptivity of the dog's behaviour. In contrast, the observer-relative perspective assumes that the problem is set by an external observer, for instance, an experimenter. From this perspective, the function of PC is to display behaviour that is appropriate with respect to an observer's assumptions about S's environment (e.g., a set of experimental conditions) and about the cognitive resources that are available to S. On this view, PC is deemed successful or unsuccessful relative to the criteria imposed by an external observer, such as an experimenter.

Wide-scoped

As a psychological explanandum, PC is subject to a range of different theories and models. Abstract descriptions of many other cognitive phenomena exhibit similar features to PC. For instance, concept learning is typically described as a computational process of inference, in which from a few experiences, a learner has to infer the concept that the experiences belong to.

Likewise, perception is sometimes understood as involving categorical decisions. An example is Harnard's (1987) explanation of people's perceptions of the similarities between two green colour shades and a yellow colour shade. Harnard finds that even if one of the green shades is closer to the yellow shade than to the other green shade on the physical wavelength spectrum, people perceive the two green shades as being more similar to each other. This task can equally be interpreted as a problem of categorising the shades on the basis of the visual input and in light of background knowledge about existing concepts or cognitive categories (e.g., GREEN and YELLOW).

The overlap of the abstract descriptions of perceptual and concept learning as well as categorisation tasks suggests that attributes associated with instances of PC overlap with attributes associated with instances of other cognitive phenomena. In this respect, the cognitive process of PC is likely to interact with other cognitive

processes (e.g., perception and concept learning). Correspondingly, a theory of PC will possibly encompass theories of other cognitive processes. It will be a wide-scoped theory.

2.2. The approach

Inspired by the observation that PC overlaps with other phenomena (e.g., perception and concept learning), the guiding research question of this thesis is: How can we study the cognitive mechanisms underlying the diversity of phenomena associated with PC? In approaching a possible answer to this question, this thesis focuses on a computational approach to PC. This approach can be positioned alongside Marr's (1982) computational level analysis of vision as an information-processing system and Anderson's (1990, 1991a) rational analysis of adaptive organisms. Both approaches have inspired the development of a reverse-engineering strategy in cognitive science. The following paragraphs provide an overview of these approaches and draw a connection to my approach.

Marr's analysis starts with identifying the problem that the system is faced with, why this problem is appropriate and what the logic of the strategy is with which this problem can be solved, instead of starting at neurophysiological details in the brain. For example, the problem of vision is to generate a 3-dimensional image from 2-dimensional input patterns on the retina. The problem is appropriate because mapping the 2-dimensional patterns onto a 3-dimensional image serves the organism to track features in its environment that are of value to its thriving and increase its chances to survive. The logic of the strategy to solve this problem is to identify properties of the visible surfaces (Marr, 1982, 36), which give rise to the structure of the world underlying the image. After having identified the problem, the analysis continues with an identification of the representations and algorithms that are appropriate to solve the problem. For example, in the case of vision, the representations are features like bars, oriented edges or blobs and properties like orientation, depth or discontinuities of visible surfaces as well as shapes and how these are organised into hierarchies that build on volumetric and surface primitives (Marr, 1982, pp. 72-73).

An example of an information-processing task is object recognition. In Marr's analysis of vision, object recognition proceeds by pairing a 3-dimensional representation from a stored collection with a symbolic index according to a set of rules. Following these rules, an object-recognition algorithm pairs novel object representations with old indices on the basis of a comparison between the novel object representation and object representations in the existing collection (Marr, 1982, ch. 5). Marr's analysis of vision ends with an investigation of the implementation of these computational processes in the neuro-visual system. This procedure presents a top-down approach to vision and contrasts with a bottom-up approach thereto, which would start from detailed investigations of the neurophysiological properties.

From Marr's perspective, all three levels are needed to explain vision but he highlights the role of the analysis at the computational level. He writes:

> The central tenet of the approach is that to understand what vision is and how it works, an understanding at only one level is insufficient. It is not enough to be able to describe the responses of single cells, nor is it enough to be able to predict locally the results of psychophysical experiments. Nor it is enough even to be able to write computer programs that perform approximately in the desired way. One has to do all these things at once and also be very aware of the additional level of explanation that I have called the level of computational theory. The recognition of the existence and importance of this level is one of the most important aspects of this approach. (Marr, 1982, p. 329)

Here, Marr expresses his belief that neither the implementational nor the algorithmic levels of analysis alone can specify the essential aspects of the task of vision, that vision is an information-processing problem. This specification happens at the computational level. The problem with the other levels is that they describe the visual processes and states as well as the activities of cells that implement these processes but they fail to explain the behaviour of the agent that sees. Explaining this is part of Marr's approach to vision, and the analysis of what vision is and why it is what it is contributes to this explanation. This is why the computational level is so important.

Anderson's (1988/2014; 1991; 1991a; 1990) rational analysis of a cognitive agent serves as a methodology to restrict the logical space of possible theories at the computational level.[3] Rational analysis is similar in spirit to the use of optimal models in evolutionary biology; both involve a rationality principle, which reflects the assumption that cognitive capacities are characterised by functions that are optimised relative to agents' needs and their environments (Anderson, 1988/2014, p. 16).

In chapter 6 of this thesis, I will argue that two assumptions that can be added to the Bayesian model can each play the role of such a principle. On the one hand, this is the weak sampling assumption in Shepard's (1987) model of generalisation and on the other hand, this is the strong sampling assumption in Tenenbaum and Griffiths' (2001) Bayesian model of concept learning. Weak sampling assumes that observed instances of a perceptual category are random samples, which just happen to fall under a concept by coincidence. Shepard suggests on this basis that the corresponding measure of whether the instances fall under a candidate concept or not is a simple measure of consistency. In contrast, strong sampling

[3] Anderson and Matessa (1990, p. 29) summarise this methodology in six steps: (1) Precisely specify the goals of the cognitive system. (2) Develop a formal model of the environment to which the system is adapted. (3) Make the minimal assumptions about computational limitations. (4) Derive the optimal behavioural function given items 1 through 3. (5) Examine the empirical literature to see if the predictions are confirmed. (6) If the predictions are off, iterate. They subsequently demonstrate the efficacy of the methodology in four case studies on memory, categorisation, causal inference and problem solving.

assumes that instances are systematic random samples and causal effects of a concept. T&G propose the size principle (the principle that concepts with a smaller size should be preferred), which I discuss in detail in chapter 6, as a measure of how well a candidate concept is justified by the given observations. I will argue in chapters 6 and 7 that the simple measure of consistency is optimised relative to inferences of perceptual categories in 'natural' environments, while the size principle is optimised relative to inferences of perceptual categories in word-learning environments. This means that in natural environments, choosing any concept that is consistent with the observations will lead to optimal (i.e., adaptive) behaviour. In word-learning environments, there are only a few optimal choices and these correspond to concepts that have a small size, which, as I will argue in chapter 6, is a function of the concept's intension. At the same time, the simple measure of consistency would be suboptimal in word-learning environments and, likewise, the size principle would be too exclusive to guarantee inferences that are adaptive in natural environments.

The point of my argument and illustrations is that both of these assumptions place different optimality constraints on the Bayesian-inference task of PC; of inferring the 'correct' concept of perceptual category given a set of observations, from these observations, how they justify the concept and some background knowledge about the concept (for details, see chapter 6). Thus, the rationality of the Bayesian approach to PC does not purely depend on the Bayesian nature of the model that I propose here, but on the additional assumptions of weak and strong sampling, which determine which choices of perceptual categories that determine behaviour are optimal, and which ones are not.

The normativity of my approach to PC is grounded in the fact that the agent's task is framed under the additional consideration of what would be the optimal behaviour of a rational agent in a given environment, where 'optimal behaviour' is behaviour that is most adaptive to a given environmental niche (cf. Anderson, 1991a), and the assumption here is that choosing the 'correct' concept produces behaviour that is adaptive. On this basis, when analysing PC behaviour from a computational-level perspective, we should look at the function that would produce behaviour that is optimal for surviving and thriving, i.e., a function that makes behaviour adaptive with respect to the environment in which the perceptual categoriser aims to thrive and survive.

Rational analyses are useful because they offer a way of analysing behaviour as rational from a 'what-if' perspective by asking how the behaviour would change under hypothetical changes in the given environment (van Rooij et al. 2018). Thus, the assumption of a rationality principle, that the cognitive function (e.g., of PC) is optimised with regards to the environment or the agent's needs, acts as a constraint on the computational level of analysis. Making this assumption allows cognitive scientists to narrow down the space of possible cognitive problems or functions that could be investigated—only those problems or functions that are optimal or adaptive should be investigated. Thus, what kind of function counts as a description of a 'rational' behaviour (i.e., rational in the sense of adaptively successful within a given environment) is important for the selection

of the descriptive framework that shall be used to investigate the behaviour. In chapter 4, I conclude that this is one potential reason for disfavouring Tversky's description of PC as a function of feature-matching—there are no obvious reasons to assume that this function is optimised with regards to the adaptive success of PC behaviour.

In recent decades, Marr's top-down approach and Anderson's rational analysis have been combined towards a broader strategy of reverse-engineering the mind. The strategy is to move through a "triumphant cascade" (Dennett, 1987, p. 227) from the computational via the algorithmic to the implementational level of analysis. Proponents of the strategy (e.g., Tenenbaum & Griffiths, 2001; Zednik & Jäkel, 2014; Zednik & Jäkel, 2016) argue that this strategy presents an optimal procedure to use behavioural data in a relevant domain of investigation and reversely inferring from that data the underlying cognitive processes that have generated the data. For example, when the cognitive process to be inferred is vision, the behavioural data from which the process is reverse inferred may be a set of recordings from a study on saccadic eye movements. What makes the procedure optimal is that it is likely to adequately describe aspects of the target phenomenon (e.g., vision) and to predict a wide variety of data associated with it (Zednik & Jäkel, 2016, p. 666). The practical implication of a reverse-engineering strategy is that cognition can be studied efficiently with a variety of tools other than neurophysiological methodology, for example, cognitive computational models (see Glossary). In following this strategy, researchers can study cognition from a reversed angle: they can infer from the observed behaviour and some rationality assumption what the most plausible candidate cognitive mechanism is that has generated this behaviour.

The strategy is often associated with Bayesian models of cognition (e.g., Chater & Oaksford, 1999, 2008; Chater & Vitányi, 2003; Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010; Griffiths, Kemp, & Tenenbaum, 2008; Tenenbaum, Kemp, Griffiths, & Goodman, 2011). These models typically analyse cognitive phenomena at the computational level. For example, a Bayesian model of vision analyses vision as a perceiver's problem to infer from an image on the retina what physical stimulus/object has caused the retinal image. The perceiver's strategy to solve the problem is Bayesian inference. The logic of this strategy is that the perceiver infers backwards, from features of the perceived image, what the features of the environment are. The strategy is rational in light of the perceiver's circumstances, under which the hidden state that has caused the input to its visual image is not directly accessible.

A possible limitation of Bayesian models of cognition is that they are idealised; they explain how an 'ideal observer' would solve a Bayesian inference task (e.g., vision), but not how she actually does this. At the computational level, the model is limited to offering a generic solution to a cognitive function or problem—the model uses a generic function to map inputs onto outputs but does not specify a set of rules for computing the function (Colombo & Seriès, 2012, p. 698). The reason for this is that many different algorithms will be able to compute the same function, which thereby underdetermines the algorithmic process. Nevertheless,

in the context of reverse-engineering, where the goal is to move through the cascade, Bayesian models of cognition have been proposed as offering constraints on possible algorithms (Zednik & Jäkel, 2014, p. 669). Likewise, Colombo and Hartmann (2017) have argued that Bayesian models in cognitive science have unifying powers, whereby they can offer fruitful constraints on the discovery, identification and confirmation of possible cognitive mechanisms in the brain (Colombo & Hartmann, 2017, pp. 466–479). In this regard, Bayesian models in cognitive science can serve as a heuristic tool to traverse from (e.g., Bayesian) investigations at the computational level to (e.g., mechanistic) investigations at lower levels in the cascade.

A detailed explanation of how Bayesian inference works with respect to PC is given in chapter 6, where I discuss Tenenbaum and Griffiths' (2001) Bayesian model of concept learning. Importantly, this model is positioned at the computational level in the cascade, where the task is specified but not how this task is actually carried out by representations and algorithms or implemented in the brain. The Bayesian unification proposed in this thesis builds on Tenenbaum and Griffiths' model at the computational level, yet, my approach is compatible with the motivation to constrain investigations at other levels in the cascade. However, the details of how exactly my proposed unification could constrain Shepard's (1987) and Tversky's (1977) competing theories at the level of representation and algorithm lie outside the scope of this thesis and form part of future research.

What is within the scope of this thesis is a discussion of Shepard's (1962, 1987) and Tenenbaum and Griffiths' (2001) models in the context of reverse-engineering. I will argue that both models exemplify this strategy—they are used as tools to reverse infer the candidate cognitive mechanisms that underlie behaviours associated with categorisation. I will borrow Machery's (2013) analysis of typical reverse inferences in cognitive neuroscience to describe the structure of these reverse inferences. Machery's original analysis is as follows.

Machery's argument structure of typical reverse inferences:

A. "When psychological process, p, is recruited by a task, pattern of brain activation, E, is likely to be found.

B. In task T, pattern of brain activation, E, was found.

C. Hence, psychological process p was recruited by task T." (Machery, 2013, p. 252)

I will use this structure to compare Shepard's and Tenenbaum and Griffiths' approaches, where this structure is instantiated with regards to observations of statistical patterns associated with behaviour, instead of patterns of neural activity. The corresponding task, T, is a Bayesian task of generalisation, the relevant patterns of behaviour are the exponential gradient of generalisation, E, and the effect of directionality, E', and the putative psychological process is a process of geometric similarity-representations, p.

2.3. Conclusion

In summary, the first section of this chapter has provided a working definition of PC, which is the ability to decide whether a perceived object belongs to a cognitive category or concept on the basis of the perceptual attributes of that object and of other objects. I have associated PC with five aspects: (1) it involves a choice among various alternative concepts and (2) the process corresponding to this choice is possibly unconscious. (3) Perceptual categories are hierarchically organised and learning them involves a problem of avoiding indeterminacy. (4) In the face of this problem, PC behaviour should be understood as adaptive. (5) A theory of PC is likely to be wide-scoped because attributes of PC behaviour overlap with attributes of other cognitive behaviours (e.g., perception and concept learning).

In the second section, I have explained the background of my approach to modelling PC. I have situated my approach in the context of the reverse-engineering strategy in cognitive science, which builds on Marr's (1982) top-down approach to information-processing systems and Anderson's (1991a) rational analysis of cognitive agents. What has been learned is that Bayesian models, such as the one that I advocate in this thesis, are positioned at Marr's computational level. At this level, we identify (a) what the problem of PC is, (b) why this problem is appropriate for the system or agent exhibiting the behaviour that we associate with PC and (c) the logic of the agent's strategy to solving the agent's problem of PC. The function of Anderson's rationality principle is to add further constraints to the analysis at this level—only problems that are adaptive or optimal should be studied. The principles of strong and weak sampling that I discuss in chapters 6 and 7 can be identified as such constraints. I have explained that the value of the reverse-engineering strategy is that it allows us to study PC with toy models of cognitive mechanisms from a reversed angle. This is useful when our access to studying PC is limited to observations of behavioural patterns whose neurophysiological underpinnings are still obscure. I have used Machery's (2013) argument scheme to describe the general structure of reverse inferences that are facilitated by this strategy. I refer to this scheme continually throughout the thesis.

In the next chapter, I begin my discussion of the two competing approaches to similarity and categorisation. I will start with Shepard's (1975; 1980; 1981/2017; 1987; 1970) geometric approach, which is also a basis for the advocated Bayesian model of PC.

Part I.

Psychological similarity

"There is nothing more basic to thought and language than our sense of similarity; our sorting of things into kinds." (Quine, 1960, p. 116)

3. Shepard's geometric approach

3.1. Introduction

My motivation to use Shepard's (1975; 1980; 1981/2017; 1987; 1970) approach to a theory of psychological similarity (PS) and generalisation as an explanans with regards to perceptual categorisation (PC) builds on an intuitive connection between Shepard's approach to generalisation and PS and the reverse-engineering approach to PC that I have explained in chapter 2. From the reverse-engineering perspective, a computational level approach to PC is helpful for explaining PC but one must also move beyond this level once the task of PC has been identified with enough precision.

Shepard's model suits this ideal; when building it, Shepard was primarily interested in unrevealing the cognitive mechanism underlying generalisation behaviour for perceptual stimuli such as colours, geometric symbols, vowel phonemes etc. His approach is a mathematical model of the 'Universal Law of Generalisation' (henceforth 'ULG'). Roughly, the ULG is an exponential function that relates a measure of generalisation to a measure of the perceived PS associated with a pair of objects. PS is a psychological measure of stimulus dissimilarity and generalisation is the objective likelihood of an agent to generalise a behaviour from one object, a, to another object, b. In Shepard's model, the function indicates an internal representation of the objects' similarity and can be studied by investigating patterns in subjects' generalisation behaviour. Because of its precision and intuitive relationship with the working definition of PC in chapter 2, Shepard's ULG is a good starting point for specifying what the function and possible mechanism of PC could look like.

The goal of this chapter is to reconstruct Shepard's proposal that generalisation is a function of geometric distance so that it can be contrasted against Tenenbaum and Griffiths' (T&G's, 2001) Bayesian theory of generalisation, which will be presented in chapter 6. The key argument in the current chapter is that PC relies on the metric axioms. Roughly, the argument is that the negative-exponential function in Shepard's model of ULG shows that geometric distance is a variable that can be experimentally investigated, and that it varies in a lawful manner. This is a case in point against popular objections against the explanatory power of PS with respect to categorisation (e.g., Goodman, 1972). Goodman had argued that PS would be too variable with respect to the context and frame of reference of a similarity-judgement task. He thought that therefore, the notion of PS would not be objective enough to satisfy scientific standards of explanation. The ULG

illustrates that a geometric conception of PS is invariant when seen in relation to a task of generalisation. One purpose of this chapter is to explain this invariance and to argue that a geometric conception of PS is a possible explanans with regards to the possible mechanisms underlying PC.

Section 3.2 outlines Shepard's approach to generalisation as a *universal law*, which he models as a monotonic function[1] that determines the relationship between PS and generalisation behaviour. Section 3.3 clarifies the relevance of the ULG for the problem of modelling PC and the method of multi-dimensional scaling. Section 3.4 outlines the theoretical assumptions of the geometric conception. Section 3.5 considers four reasons for why PS should be constrained by the metric axioms. Section 3.6 concludes with a connection between Shepard's model of the ULG and the problem of modelling PC.

3.2. Shepard's Universal Law of Generalisation

The ULG is a psychological law that governs generalisation behaviour. The law models generalisation as a psychological phenomenon that describes a pattern of behaviour that is invariant under changing conditions (Shepard, 1987, p. 1318)[2]. In light of the ULG, generalisation behaves in the same way when considering varying species, stimuli and modalities (cf. figure 3.1). For instance, according to the model, the shape of generalisation is the same when comparing how humans treat different geometric shapes and how pigeons treat different colour shades (cf. subfigures A and E).

Mathematically, the ULG is a negative exponential curve that maps the perceptual similarity of two objects onto subjects' likelihoods to generalise their behaviour from one to the other object. Thus, when knowing how perceptually similar two objects are, one can use the function to predict how likely subjects

[1] A monotonic function is a mapping between ordered sets that maintains the order of the sets. For example, if a function between two ordered sets of points is monotonically decreasing, then the order between the sets never increases in the mapping. The ULG is an example of a monotonic function because it maintains the decreasing order of two sets of stimuli. This order is typically represented in the row and column of a similarity matrix. An example for such a matrix is presented on the left of figure 3.2.

[2] On Hempel's 1942 classical account, laws are regularities that meet certain further conditions. These conditions help to make sure that the regularities are non-accidental. Hempel illustrates the distinction between accidental and non-accidental conditions with the example of the Greensbury School Board for 1964, in which all members happen to be bald. Hempel's intuition is that there is not a law of baldness that governs the fact that all members of the Greensbury School Board for 1964 are bald. In contrast, the fact that all gases expand when heated under constant pressure is a regularity that is non-accidental, and is therefore, following Hempel, a law. One example for a relevant condition is that laws must be general: laws are regularities associated with phenomena that are invariant when repeated under a range of different circumstances in nature. Thus, if there was a psychological law that adequately described generalisation behaviour, then it could be expected that generalisation behaviour expresses patterns that stay invariant across a range of different circumstances in nature.

will be to generalise behaviour from one object to another. This is exactly what Shepard did; he used a geometric model of PS and predicted on this basis the pattern of generalisation that is described by the negative exponential curve in figure 3.1. To predict patterns of generalisation on the basis of PS, Shepard used a measure of geometric distance. To obtain the ULG, Shepard determines the particular distances in PS space such that how likely it is that the model generalises from one object to another is a monotonic function of the corresponding points' geometric distance in the PS space. Shepard's model predicts generalisation likelihoods best when this function has a negative exponential shape. What the ULG says can be summarised by the following proposition.

Definition 3.2.1 (The universal law of generalisation)**.** For any pair of stimuli, i and j, the empirical probability, g_{ij}, of an organism to generalise a response from i to j is a monotonic function of the psychological distance between i and j, d_{ij}, with a negative exponential shape.

Figure 3.1 illustrates that the law models a variety of case studies of generalisation. Subfigures E and H depict the relation between pigeons' generalisation behaviour and a physical measure of stimulus distance. For example, the physical difference between colours is measured by their differences in physical wavelengths of light. Subfigures B and K depict the relation between humans' generalisation behaviour and a psychological measure of stimulus distance. A measure of psychological difference in colour is how similar or different colours are perceived to be. This measure relies on psychophysical scaling tools, such as tests of how well (e.g., quickly or accurately) people detect or discriminate between different colours. Subfigure L depicts data on humans' categorisations of Morse Code signals. In this data set and according to Shepard's geometric-distance model, the psychological difference between Morse Code signals is a function of how much these signals overlap in their length (number of components) and style (dots versus dashes). In all cases, the function that maps stimulus distance in PS space onto probability of generalisation has a negative exponential shape, no matter what species the agent belongs to nor what stimuli it generalises. This is the sense in which the function describes a universal gradient of generalisation.

The following paragraphs provide two examples of what type of entities the ULG relates to each other. These examples connect what has been said about the ULG to the phenomenon of PC. Example A illustrates that behaviour associated with PC is often studied in the form of a generalisation task (e.g., 'if a is a 'fep' is b a 'fep' as well?'). One explanation, which is illustrated in the examples, is that the agent's decision as to whether or not the agent should generalise depends on the underlying perceptual categories. Typically, a representation of a perceptual category (cf. Glossary) is more abstract or comprehensive, than a representation of the PS between two individual objects in a similarity-judgement task. However, if understood as a psychological representation, a perceptual category or concept is also defined in terms of the PS amongst its possible members. It is generally unclear what the exact relation between the PS between two objects and their

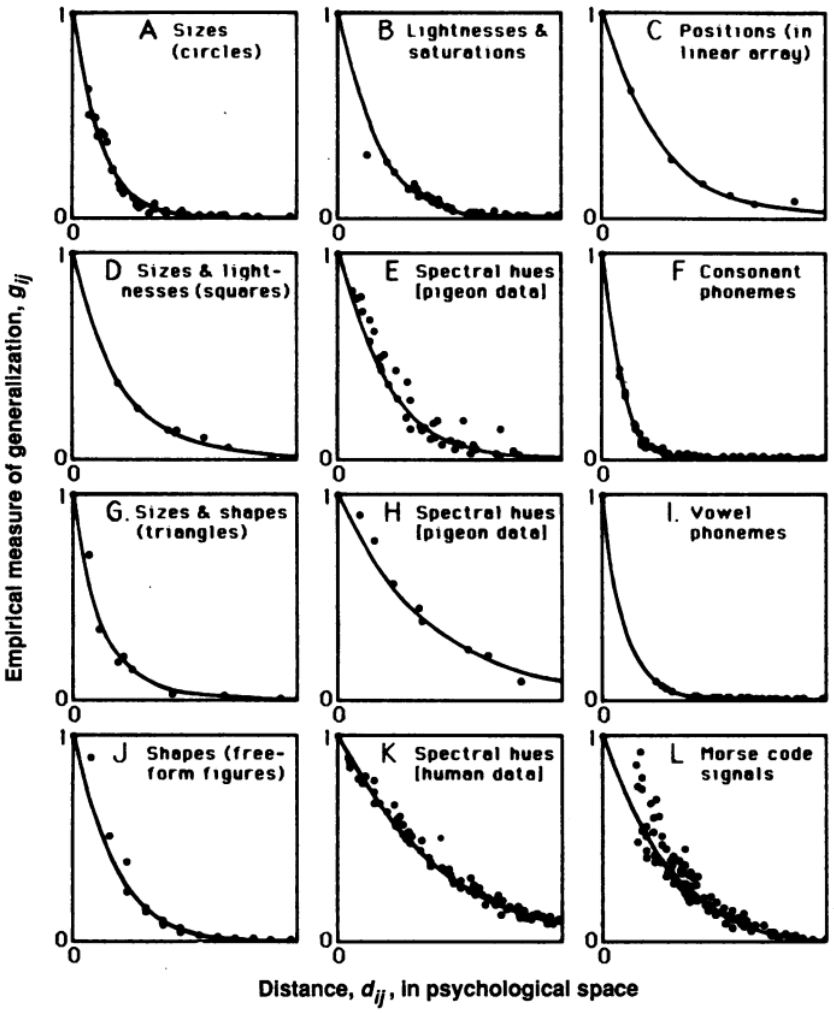


Figure 3.1.: A graphical illustration of several instantiations of the universal law of generalisation. Each subfigure, A-L, represents a negative exponential function that maps stimulus distance or dissimilarity (on the x-axis) monotonically onto generalisation probability or confusion likelihood (on the y-axis). From "Toward a universal law of generalization for psychological science," by Shepard, Science, 237(4820), p. 1318. Copyright 1987 by The American Association for the Advancement of Science. Reprinted with permission.

belongingness to a concept is. Thus, I take it that, practically, the data sets in figure 3.1 could be interpreted as evidence for the hypothesis that a process of PC is happening. The following two examples illustrate this point.

(A) People's categorisation of circles

One example of the ULG is McGuire's (1961) study of how people categorise circles according to their relative sizes. Subjects were trained with sets of black circles that varied in their diameter sizes. In the first condition, sizes varied in big

steps of .14cm. In the second condition, sizes varied in small steps of .07cm. Each circle was assigned a numerical label. The smallest circle was labelled with a '1' and the largest circle was labelled with a '9'.[3] In the test phase, subjects were asked to assign one of the learned labels to a novel circle. McGuire's hypothesis was that it is easier to assign the correct label to circles in the first condition than in the second condition. A label was correct if the experimenter had assigned it to a circle of the same size in the test phase.[4] His reasoning was that circles in the first (big) condition are less similar to each other than circles in the second (small) condition. Hence, it should be easier to discriminate between them.

McGuire's (1961) results (see graph A in figure 3.1) illustrate that the ULG is a good description of how people categorise circles according to their relative sizes. With an increase in the relative similarity between two circles (with respect to their diameter sizes), there is an exponential increase in people's likeness to treat (i.e. label) them as the same. In other words, people were more likely to give two circles of a similar diameter size a similar label, and they were less likely to give circles with different diameter sizes the same label. Thus, knowing the similarities between conditioned pairs of circles, McGuire was able to predict how people will categorise circles they had not yet seen.

(B) Pigeons' categorisation of colours

Another example is Guttman and Kalish's (1956b, henceforth 'G&K') study of pigeons' pecking responses to a key that was illuminated in different colours. In the training phase, G&K conditioned pigeons with keys of colours within some range on the physical wavelength spectrum (e.g., green colours). In particular, the conditioned stimuli were wavelengths of 530, 550, 580 and 600 $M\mu$. Upon pressing the key, a door would open the way to a food magazine. When pigeons pecked at the key when it was lit with wavelengths that were outside that range (e.g., red colours), no food would be offered. In the test phase, a wavelength of light (e.g., $530M\mu$, corresponding to green light) was projected onto the key. The overall range of wavelengths of the test stimuli was from $460M\mu$ to $660M\mu$ in $10M\mu$ intervals for each of the conditioned stimuli types, and pigeons had to respond with several pecks to open the magazine when the door was lit with the right colour (i.e., within the range $530 - 600M\mu$).

G&K measured the degree of a pigeon's generalisation as the number of pecks that that pigeon would give a key relative to all pecks from that pigeon across

[3] For instance, in the first condition, a circle with a radius of .7 cm would be labelled with a '1', a circle with a radius of .21 cm would be labelled with a '2', a circle with a radius of .35 cm would be labelled with a '3' and so on. Likewise, in the second condition, a circle of size .37 cm would receive a '1', a circle with size .44 cm would be labelled '2' and so on.

[4] For instance, if in the training phase of condition I a circle of size .21 cm was paired with a '2', and a circle of size .35 cm was presented in the test phase, then the correct response would be a '3'. Likewise, if in the training phase of condition II a circle of size .44 cm was paired with a '2', and a circle of size 51 cm was presented in the test phase, then the correct response would be a '3'.

trials. Generalisation was considered to be more likely if pigeons pecked on average relatively more often. For instance, pigeons in the $530M\mu$ group should be more likely to peck in the test phase if the key was projected with $540M\mu$ than if it was projected with $550M\mu$ or $600M\mu$. G&K's hypothesis was that pigeons will be more likely to peck if they expect food in the magazine and whether they expected food depends on the similarity between the conditioned colours and the unconditioned/test colours. G&K's results confirm these predictions. Graph E in figure 3.1 shows that the correlation between the recorded wavelengths and response likelihoods takes the shape of a negative exponential curve. Together, examples A and B illustrate that there is a correlation between similarity and generalisation.

On reflection, there are two issues with McGuire's and G&K's models. Firstly, the models do not seem to predict aspects of the cognitive mechanisms involved in PC. This is because they use physically objective measures of behaviour and stimulus properties. In McGuire's study, the objective measure of stimulus dissimilarity is their difference in physical diameter size and the measure of generalisation is people's objective likelihood to correctly label a circle. In G&K's study, the objective measure of stimulus dissimilarity is the difference in physical wavelength of illumination and pigeon's tendency to generalise is measured by the objective number of pecks that they give the key. None of these measures is a psychological measure. In contrast, Shepard's geometric-distance measure (on the x-axis in figure 3.1) is supposed to be a measure of the unconscious perceptual experience associated with subjects' generalisation behaviour. Thus, from these studies alone, there is no reason to believe that subjects in examples A and B in fact generalise because of how similar or different the stimuli look to them. Secondly, without regards to the negative exponential curve in figure 3.1, there is no link between the results of these two studies on PC. This is problematic because it makes it difficult to integrate the results of these studies into a theory of generalisation or PC.

In contrast, the ULG relates these models elegantly by predicting the results of examples A and B simultaneously. It also adds information to the interpretation of McGuire's and G&K's findings. The additional information is that the data is not only a mere correlation of objectively observable events but is caused by mental states of the agent. In particular, on Shepard's approach to generalisation and PS, the ULG (exponential function in figure 3.1) describes a correlation between the physical similarity of the stimuli and generalisation behaviour. But it also describes how generalisation should change with a change in the PS associated with the stimuli (a representation in the agent's mind). The prediction that generalisation should be relatively improbable under relatively dissimilar pairs of stimuli is justified by the additional assumption that agents represent the psychological properties (e.g., the similarity) of objects in their mind and that the accuracy of this representation, and the generalisation response plays a role for the agent's success (e.g., their adaptive success). Thus, the ULG contributes with a method of predicting McGuire's and G&G's data from a psychological perspective.

Taken together, this section has explained that the ULG presents a rationale for why the exponential gradient represents a cognitive phenomenon and not a mere stimulus-response behaviour. With regards to PC, what allows the ULG to claim the status of a psychological law is the assumption that a subject's tendency to categorise two stimuli as the same (or different) depends on the degree of PS between the stimuli (i.e., on an internal representation thereof), instead of physical stimulus difference. The next section explains why the ULG matters for an explanation of PC. It presents two ingredients that motivate ULG for this purpose: (1) A theoretical strategy of reverse-inference and (2) an empirical method of predicting generalisation behaviour with a psychological measure of stimulus difference.

3.3. The universal law and perceptual categorisation

3.3.1. Reverse inference

The following paragraphs explain how Shepard's (1987) model of the ULG resembles a reverse-engineering approach to PS and generalisation. The target of Shepard's model of generalisation was to find the psychological function that relates a pair of stimuli and the agent's response in a generalisation task. The available evidence for inferring this function was limited; it consists in the available recordings of subjects' explicit judgements about how similar the stimuli are or the average probability of subjects to confuse these objects. To carry out this inference, Shepard relies on various additional assumptions that are more explicitly reflected in his earlier work (especially in Shepard (1962, 1981/2017)).

In this work, Shepard claims that there exists some structural correspondence between subjects' behaviour and their internal mental representations of categories[5]. Shepard's motivation for this claim was that subjects' behaviour is implicitly governed by evolutionary norms; Shepard had though that the most adaptively successful behaviour is such that it indirectly reflects aspects of natural kinds (e.g., whoever avoids mushrooms of the poisonous kind and eats those of the edible kind gets the representations right). In light of the structural-correspondence claim, Shepard inferred that the observed exponential gradient of generalisation behaviour reflects subjects' 'internal' preferences to generalise the same behavioural consequence towards objects that are more similar to each other (i.e., in the model, points that are closer in PS space) than to objects that are dissimilar (i.e., far apart in the model). Intuitively, this is the behaviour that is most probable to ensure the subject to be adaptively successful (e.g., it would be a 'miracle' that an agent could survive if there was no such mapping between the internal structure and the behaviour, such as eating or avoiding).

[5]In his 1981/2017, he calls this correspondence a 'second-order isomorphism'.

On the basis of these considerations about structural correspondence and adaptive success, Shepard inferred that the observed statistical structure of subjects' behaviour (e.g., the observed confusion probabilities) reflects something about what the structure of subjects' internal spaces of concepts look like. Therefore, he believed that the study of the objective generalisation probabilities associated with subjects' behaviour would help him to reverse-infer from the subjects' behaviour what the structure of subjects' internal representations of categories looks like.

The structure of Shepard's (1987) scientific inference can be simplified with Machery's (2013) reverse-inference scheme that I have discussed in chapter 1.

Argument structure of Shepard's approach:

A. When psychological process, PS, is recruited by a generalisation task, an exponential gradient of generalisation, E, is likely to be found.

B. In generalisation task T, E, was found.

C. Hence, PS was recruited by T.

In Shepard's model, PS refers to a cognitive process that uses a representation of PS to generate an appropriate response to a stimulus on the basis of aspects of another stimulus.

The assumption that PS drives generalisation and, as I claim here, the mechanism of PC, connects Shepard's theory of generalisation to the problem of modelling PC. In Shepard's theory, generalisation is the cognitive act of deciding whether i and j, despite the fact that they are distinct perceptual objects, belong to the same natural kind or concept (Shepard, 1987, p. 1319). Assuming that a concept in Shepard's theory is equivalent to a cognitive category, the cognitive process that underlies generalisation in Shepard's model is an instance of PC. In support of this point, S. A. Frank (2018, p. 9803) argues that "'[g]eneralization' arises because perceived similarity may describe recognition of a general category. For example, two circles may have different sizes, colors, and shadings. Perceived similarity arises from the generalized perception of 'circle' as a category." Generalisation from i to j on the basis of their perceived similarity should be interpreted as the decision that i and j belong to the same category. Thus, given the model's assumption that PS drives generalisation, deciding whether i and j belong to the same perceptual category is driven by PS as well.

This makes the connection between Shepard's model of ULG and a reverse-engineering approach to PC more explicit. The ULG figures into the scheme of reverse-engineering because the geometric-distance model is only a means to finding out what the assumed perceptual categories or concepts proposed in the model look like, and how they can be eventually understood as efficating generalisation behaviour. Let us take a closer look at how ULG can be understood to reverse-infer PS.

3.3.2. Multi-dimensional scaling

As argued in section 3.2, to predict patterns of generalisation on the basis of PS, it is necessary to have a measure of the PS between the objects. Roughly, Shepard's solution to this problem is to measure PS as geometric distance. He assumes that there exists a PS space with multiple dimensions in which points represent perceptual objects (e.g., Munsell colour chips or Morse Code signals) and the distances between these points represent the relative dissimilarity between the objects. The model of geometric PS space represents aspects of the agent's mental states. For example, the mental image of a red colour shade is a point in a space in which the dimensions are hue, saturation and brightness. Shepard defines PS as the psychophysical function that maps between properties of the physical stimuli and an appropriate behavioural generalisation response (cf. Shepard, 1981/2017). This introduces the particular method that Shepard uses to reverse-infer PS from patterns of generalisation data—the method of multi-dimensional scaling (henceforth 'MDS').

An MDS algorithm transforms empirical recordings of generalisation data such as similarity judgements and confusion probabilities into a distribution of points in a multi-dimensional, geometric, space. The goal of MDS is to find the cognitive process and representation of the objects that accurately generates the generalisation data. In other words, this is a form of reverse-engineering the cognitive process that has generated the data. The MDS algorithm uses information about metric distances between points to generate a function that accurately predicts this generalisation data.[6] In calculating distances, MDS 'scales' the data and thereby offers a psychological interpretation of stimulus difference. This transformation of a similarity judgement or confusion probability to a point in geometric space can be interpreted as a perceptual representation of how similar two stimuli are. In the model, this representation corresponds to a "mental arithmetic that mimics the distance formula" (Borg & Groenen, 2005, p. 3). In other words, geometric distances in the model represent subjects' perceptions of similarities between objects and the multi-dimensional space models a PS space. Recently, this model has been taken a step further towards the explicit claim that the dimensions of geometric space represent qualities associated with subjects' perceptual experiences, for instance, the perceptual experience of colour (cf. Gärdenfors, 2000, pp. 6-15). This characterisation of MDS can be summarised in the following proposition.

Lemma 1 (Geometric Similarity)**.** If two perceptual objects, i and j, can be represented as vectors in a psychological similarity space that is structured by geometric quality dimensions, then how perceptually similar i and j are is expressed

[6] Computationally, the most important aspect of the method is that the distribution of points in multi-dimensional space is modelled along as few dimensions as possible, until there is no better fit with the linear differences in the ordinal data with n dimensions than there would be with $n+1$ dimensions. The ergodicity of the distribution ensures that the resulting configuration is objective, in the sense that any starting point of modelling the data will result in the same spatial configuration of points.

in the geometric distance between their corresponding vectors in psychological similarity space.

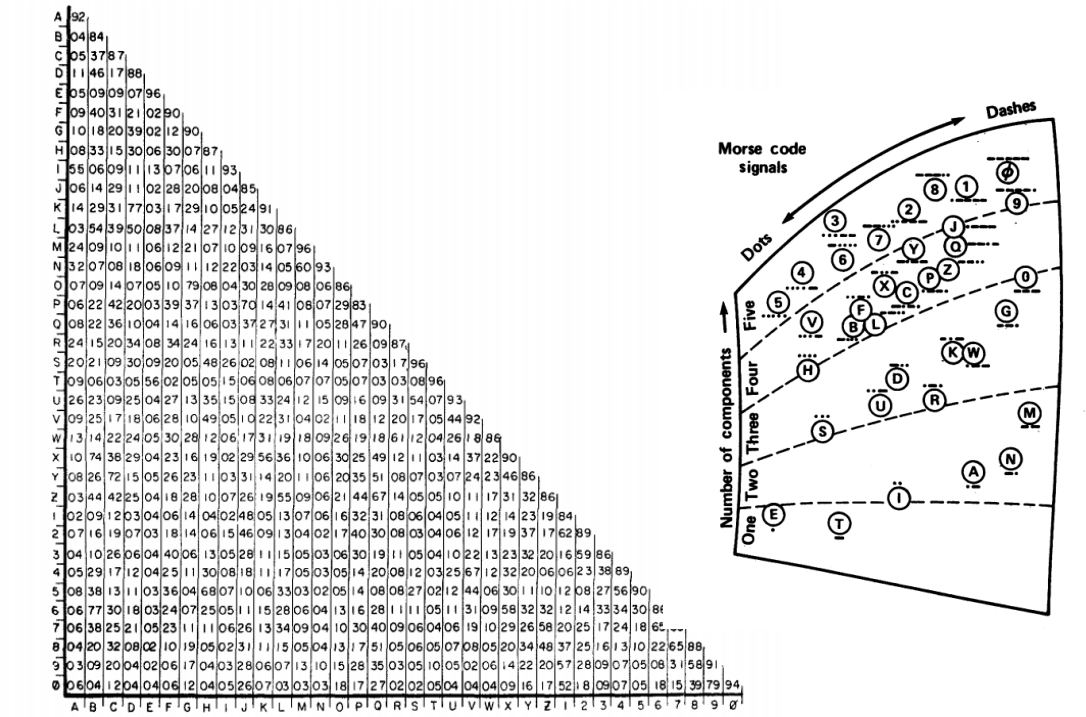


Figure 3.2.: On the left: Similarity matrix indicating judgements of similarities of Morse Code Signals. From "A measure of stimulus similarity and errors in some paired-associate learning tasks," by Rothkopf, 1957, Journal of Experimental Psychology, 53(2), p. 97. Copyright permission is in the public domain. Indexes indicate averages of subjects' confusion probabilities associated with a pair of signals. On the right: a 2-dimensional configuration of Rothkopf's data. Morse code signals are classified according to the number of components on the y-axis and the relative proportion of dots versus dashes on the x-axis. For example, the signal for letter 'V' has 4 absolute components, and a relative proportion of 3 dots to 1 dash. From "Multidimensional scaling, tree-fitting, and clustering," by Shepard, 1980, Science, 210(4468), p. 391. Copyright 1980 by The American Association for the Advancement of Science. Reprinted with permission.

An example of MDS is Shepard's (originally 1963, as cited in Shepard 1980, p. 391) analysis of Rothkopf's (1957) Morse Code data. The left of figure 3.2 presents Rothkopf's original 36×36 similarity matrix, which indexes the ordinal differences between subjects' average confusion probabilities associated with each pair of Morse code signals in the stimulus set of Rothkopf's study.[7] The right

[7]Rothkopf computed these indexes by taking the confusion probability associated with any judgement for each pair, e.g. for {A,B}, and its respective counterpart, e.g. {B,A}, adding the results and dividing by two. Thus, where originally there was a '09' at position [∅,A] and a '03' at position [A, ∅], the new matrix indicates a '06' at position [∅, A] because $09 + 03 = 12, 12/2 = 06$. Thus, the matrix only shows the values for an idealised pair of signals.

of figure 3.2 illustrates the results of Shepard's transformation of Rothkopf's data into a 2-D configuration of points in multi-dimensional space (as cited in Shepard 1980, p. 391). Shepard models the signals along two dimensions. The y-axis indicates the absolute number of components that each signal has, where this value ranges from 1 to 5 components. For example, the signal '...-', which stands for the letter 'V' in the International Morse Code alphabet, has 4 absolute components. The x-axis indicates the relative number of dots and dashes that is associated with each signal. For instance, the signal '...-' ('V') has a proportion of 3 dots against one dash.

Taken together, this section has connected Shepard's method of MDS to the reverse-inference scheme suggested in chapter 1. This method operates under the assumption that the cognitive mechanism that generates patterns of generalisation behaviour operates on representations of perceived similarities between the relevant stimuli. The next section explains more precisely the proposal that PS can be measured as a geometric distance function.

3.4. Assumptions of the geometric approach

The assumption that the structure of perceptual categories or concepts is a function of geometric distance deserves special attention because in Shepard's model, the ULG is derived from this assumption. (In chapter 6, it will be shown that some aspects of the ULG can also be preserved without the geometric-distance assumption.) This assumption guides the scientific inference that the observed patterns of generalisation are based on a process inside the agent's mind that involves mental representations of relations between the relevant objects. This is why the ULG is a law about psychological mechanisms, as contrasted earlier with correlations of physical events (e.g., objective probability of generalisation and wavelength of light). Two aspects of this assumption should be highlighted. The first aspect is that there are multiple ways to measure geometric distance, suggesting that there might not only be one unique function that could accurately model PS. The second aspect is the reliance of the geometric distance model on the metric axioms. To start with the first aspect, the following paragraphs explain two prominent distance functions.

The first function is the city-block distance, which measures the distance between a pair of points by taking the sum of their distances along each axis in a geometric space. Formally, this is expressed by the following equation.

$$d_{ij} = \sum_{i=1}^{n} |x_i - x_j|, \tag{3.1}$$

where d_{ij} stands for the psychological dissimilarity (i.e., the geometric distance) between a stimulus i and a stimulus j. The formula says that for a pair of

points, the dissimilarity between their associated objects is the sum of the points' distances along each axis in geometric space.

The second function is the Euclidean distance, which measures distance along a direct path by taking the diagonal between any pair of points or vectors in geometric space. More formally:

$$d_{ij} = \Big(\sum_{i=1}^{n} |x_i - x_j|^2 \Big)^{1/2} \tag{3.2}$$

In equation 3.2, the dissimilarity between two points is their squared distance, raised to the power of 1/2.

Note that in Shepard's (1987) model, dissimilarity is interpreted to mean the opposite of PS. Correspondingly, PS is the inverse of geometric distance so that distance corresponds to dissimilarity, $d(\cdot, \cdot)$. Hence, according to the model, the closer two objects are in PS space, the more similar they are. Conversely, the greater the distance associated with two objects is, the less similar they are. More formally, let a and b each represent a vector in multidimensional space, and their geometric distance be represented as the function $d(a, b) = 1/s(a, b)$. Figure 3.1 suggests that the distance function has a negative exponential form with respect to generalisation tasks.

The method of MDS uses either of these distance measures to derive the ULG. It has been suggested that which distance measure is more appropriate for modelling PS depends on the context, which can be specified as differences in the structure and relations between the dimensions. However, there is no consensus about which measure is better in which context. Shepard (1980, p. 394) argues that the measures obtain different accuracies depending on whether the stimuli are "perceptually unitary" or whether they are "analyzable". Homogeneous colours are an example for perceptually unitary stimuli and differences between them should be measured with the city-block distance. Geometric shapes are examples for analysable stimuli and should be measured with the Euclidean metric. Gärdenfors (2000, p. 24) argues for a similar difference between "integral" and "separable" dimensions. Gärdenfors' example for integral dimensions are the dimensions of hue, saturation and brightness. He argues that differences in colours should be measured with the Euclidean distance instead. Separable dimensions, such as *size* versus *shape*, should be measured with the city-block distance. Thus, both distance measures of PS may be more or less appropriate, depending on the given context.

3.4.1. The metric axioms

Both measures satisfy the most important theoretical assumption in the geometric model; that PS is governed by the geometric axioms—a geometric distance function, δ, is a mapping from a pair of points in a geometric space to a non-negative number, under the condition that δ maps numbers onto pairs of points

in a way that satisfies the metric axioms. Geometric distance is constrained by the metric axioms: symmetry, minimality and triangle inequality.

1. **Minimality**: $\delta(a, b) \geq \delta(a, a) = 0$.

   In words: The distance between an object and itself is equal for all objects, and must be smaller or equal to the distance between two objects. It is only equal to the distance between two objects if and only if these are identical, that is, if and only if $a = b$.

2. **Symmetry**: $\delta(a, b) = \delta(b, a)$.

   In words: The distance between two objects must be the same, regardless of the direction from which their distance is measured. The distance from a to b must be the same as the distance from b to a.

3. **Triangle inequality**: $\delta(a, b) + \delta(b, c) \geq \delta(a, c)$.

   In words: In a triadic comparison, one distance must always be shorter or equal to the sum of the other two distances.

The axioms lead to the following constraints on possible PS processes.

1. The PS between an object and itself is constant, and must be greater or equal to the PS between two distinct objects.
2. The PS between a and b must be the same as the PS between b and a.
3. Given three objects, a, b, c, the PS of any pair, (a, c) must be shorter than the PS of any two pairs, (a, b) and (b, c), together.

In summary, this section has presented the geometric conception of PS. The next section presents three arguments in favour of this conception.

3.5. In favour of the geometric conception

Empirical support

A first argument in favour of the geometric conception focuses on the successful application of the method of MDS. At Shepard's time, this method achieved more empirical support than earlier attempts to interpret similarity-judgement data. One of such attempts was a linear model that only relied on the ordinal data from subjects' similarity judgements or their confusion probabilities. On an ordinal scale, the numerical distances have a linear relationship to each other. This method is less predictively powerful; it cannot derive the ULG as it predicts a linear instead of an exponential relationship between the probability of a generalisation responses and stimulus similarity.

Another problem of ordinal analyses is that they complicate the empirical support for the hypothesis that there is a unique cognitive mechanism that corresponds

to behaviour that is associated with these similarity judgements. Linear distributions associated with categorisation data vary depending on whether the data is from studies with one or another species and one or another type of stimulus. McGuire's and G&K's results illustrate this: In both cases, probability of generalisation decreases with physical stimulus dissimilarity but the way it decreases varies from A to B. Therefore, the data from these studies cannot be regarded as instances of support for a common theory of PC. In contrast, the ULG accumulates empirical support across different species and stimuli and is a better candidate for such a theory. Given that the ULG is derived with MDS, MDS receives indirectly more empirical support than the ordinal model.

The assumption that generalisation depends on geometric distance has also motivated further empirical research that lends indirect support for the ULG from the empirical sciences over recent decades. From their comprehensive survey and analysis of more recent ethological generalisation data, Ghirlanda and Enquist (2003, p. 15) conclude that there are aspects of generalisation that are universal: "patterns of generalization are largely independent of systematic group (evidence is available for insects, fish, amphibians, reptiles, birds and mammals, including humans), behavioural context (feeding, drinking, courting, etc.), sensory modality (light, sound, etc.) and of whether reaction to stimuli is learned or genetically inherited". They note that "[...] such gradients are better described by Gaussian curves than by exponentials" (ibid.). However, mathematically, the Gaussian form is a special case of the exponential form of the gradient when a multi-dimensional psychological space is assumed (S. A. Frank, 2018). Overall, this research shows that the geometric conception has suggested further questions about the shape of the generalisation function and that these questions have afforded more recent practical applications and additional empirical support for the ULG.

Visual transparency

A second argument in favour of the geometric conception is that it allows for more simplicity in the interpretation of the behavioural data. One way in which this can be seen is by looking at the MDS method. Figure 3.2 illustrates this. The MDS solution makes the overall relationship between the averaged PS data explicit and visually accessible. The resulting configuration on the right simplifies the matrix on the left in this sense by rearranging the data points from the left into a 2-dimensional configuration on the right. The simplification makes the data also more explicit. For example, the information revealed by indices on the left is restricted to the pair-wise comparisons of data but the spatial configuration on the right visualises the implicit psychological structure across these comparisons. What is more, the spatial configuration allows us to classify the Morse Code signals into five categories, as illustrated by the band that is cut across the 2D space. This illustrates that the visualisation of similarity-judgement or confusion data with MDS reveals patterns that were originally obscured.

Informativeness

A third argument for the spatial solution is that it makes the data more comprehensive. In the Morse Code example from figure 3.2, the spatial configuration reveals how the original pair-wise PS judgements relate to each other across the whole signal system. For example, the band-classification explains why subjects are more likely to wrongly judge Morse Code signals of the form '_._' (representing the letter 'X') and '._ _.' (representing the letter 'P') to be the same; these signals lie in the same band in proximity space, while other signals lie in different bands. In other words, the band pattern allows one to infer where in the distribution subjects may place category boundaries. Thus, the spatial configuration is useful to interpret the original data in novel ways.

When considered in the context of a reverse-engineering approach to PC, the geometric approach is useful in serving an 'as-if' explanation of how PC could work.[8] The idea is that the MDS algorithm simulates the mental process that is responsible for computing similarities in the subject's mind. Roughly, the process takes a measure of PS between two objects, a and b, and generates a probability of generalising from a to b, given this similarity measure. Changes in generalisation behaviour can then be explained by reference to a psychological mechanism that works 'as if' subjects compute geometric distances between perceptual object representations in their internal psychological spaces.

3.6. Conclusion

In summary, this chapter has reconstructed Shepard's (1987) proposal that PS is a geometric-distance function (section 3.3.2). Roughly, two objects are relatively similar if the distance of their corresponding points in psychological similarity space is relatively small. In Shepard's geometric model, the objects are perceptual kinds and the dimensions represent perceptual attributes (e.g., for colour, these are the hue, saturation or brightness dimensions). The use of Shepard's geometric model of PS is to derive the ULG, which is Shepard's cognitive approach to generalisation. I have explained what the ULG is (section 3.2): a psychological law that states that the probability to generalise from one object to another is a negative exponential function of the psychological distance or dissimilarity of the objects. Objects that are closer in this space obtain a higher probability of being confused or judged to belong to the same category. I have discussed the ULG in the context of two case studies: peoples' categorisations of circles and pigeons' categorisation of colour.

I have argued that Shepard's ULG can be understood as a psychological law of PC behaviour and positioned this view on ULG in the context of a reverse-engineering approach to PC (section 3.3). The method of MDS is a key step in

[8] For an excellent explanation of how 'as-if' explanations are used in models that analyse cognitive behaviour as rational, see van Rooij et al. (2018).

this relation. This method offers an approach to modelling PS representations and algorithms, thereby moving beyond a computational level approach to PC. I have subsequently favoured the geometric conception of PS based on three arguments (section 3.5). Firstly, the method of MDS obtains more empirical support than ordinal methods. Secondly, the method makes the generalisation data visually accessible. Thirdly, the geometric conception allows for novel interpretations of the data. Lastly, I have explained the model's key assumption that PS relies on a geometric-distance function and on the metric axioms (i.e., minimality, symmetry and triangle inequality, section 3.4.1).

Overall, the key argument put forward was that PC is a function of geometric distance and relies on the geometric axioms. The key support for this argument was the intimate connection between Shepard's ULG, which he had derived from a metric PS spaces model, and the phenomenon of PC. Both phenomena reflect a cognitive function of categorising perceptual objects. Shepard's ULG is mathematically elegant and has a broad scope; it predicts generalisation behaviour across species, stimulus domains and modalities (figure 3.1).

I close this chapter with a critical reflection on two problems with Shepard's ULG that are worth mentioning about his approach. The first problem is that in Shepard's (Shepard, 1987) examples, ULG is typically derived from data that is aggregated across subjects. For example, in the ordinal matrix in Rothkopf's (1957), each index reflects the average of a group of subjects' judgements of the similarity associated with a stimulus pair and does not directly reflect subjects' individual judgements. The worry is that ULG may not reflect information about a possible mechanism that may have generated the data patterns in any single subject. In other words, Shepard's ULG may not be a description of the function that explains generalisation behaviour in an individual subject. However, it should be noted that this criticism possibly concerns only the result of ULG and not the practical viability of the MDS method. In principle, MDS can be used to construct configurations of individual data (Borg & Groenen, 2005, ch. 21).

The second problem concerns the assumption that PS relies on the metric axioms, and in particular, on the axiom of symmetry (section 3.4.1). The assumption is problematic in two respects. In one respect, the assumption limits the power of the geometric model to account for some effects in the generalisation data. For example, in figure 3.2, the spatial solution on the right lacks information about ordering effects that were initially present in the original data matrix on the left. For example, the original matrix in figure 3.2 indexes different values for the two distinct pairs (A,B) and (B,A) but the transformed matrix is more restricted; it indexes an average value to represent these two pairs indistinctly (i.e., A is as similar to B as B is to A). Thus, with respect to ordering effects, the spatial configuration is useful to reveal some aspects of the generalisation data (e.g., classification patterns) but it is not a perfect solution for revealing all aspects of it.

The second respect is philosophical and concerns a problem of directionality. Shepard (1981/2017) sees PS as a mapping from processes in the world (e.g.,

physical relations between two stimuli) to mental states (e.g., mental representations of their similarities). Given that he defines PS as a geometric-distance function, this implies that the world-mind relation must be symmetric. However, there are strong reasons to doubt that this relationship is symmetric. From a philosophical perspective, the relation between processes in the world and mental processes is inherently directional because PS states are *about* states in the world (see A. M. Isaac, 2013, for discussion). A possible argument against the geometric model is then that it lacks an account of aboutness because the model assumes that PS is symmetric. To account for the aboutness of mental states (e.g., representations of similarity), one needs to account for the directionality in the function that relates PS states to states of objects and the relations between them.

These respects are distinct: the first sense of directionality is present in the empirical data on similarity judgements but the second sense of directionality concerns the aboutness of mental states. The first sense is an objectively measurable fact while the second sense is an assumption of a theory of the mind. The next chapter focuses on the empirical problems of directionality and discusses a possible alternative approach to PC. This approach centres on Tversky's set-theoretic model of PS.

4. Tversky's feature-matching approach

4.1. Introduction

This chapter discusses Tversky's (1977) theory of PS as a set-theoretic function of matching features. The chapter contrasts Tversky's feature-matching model of PS with Shepard's (1987) geometric model of PS. I focus on Tversky's explanation of the directionality associated with empirical ordering effects in similarity judgements. An example for this directionality is the finding that people typically judge Tel Aviv to be more similar to New York than New York to Tel Aviv. I argue that Tversky's account of why directionality occurs contributes to an understanding of why patterns of PC sometimes change depending on the context. An example of this context-sensitivity is that sometimes, it is more natural to group the first pair in figure 4.1 with the second pair, while other times, it is more natural to keep these sets of pairs separate. This is interesting because explaining effects of directionality and context-sensitivity is difficult to do with the ULG and the geometric conception of PS. The ULG describes a regularity (i.e., it is context-insensitive) and the geometric conception relies on the axiom of symmetry.

Section 4.2 clarifies the key assumptions of the feature-matching model. Section 4.3 motivates Tversky's (1977) feature-matching theory of PS as an alternative to Shepard's (1987) geometric theory of PS on the basis of the finding of directionality. Section 4.4 illustrates how Tversky's model accommodates asymmetries with a case study of Rothkopf's Morse Code data. Section 4.5 explains how the feature-matching model, in conjunction with Tversky's diagnosticity principle, can account for effects of context-sensitivity in PC. Section 4.6 evaluates the feature-matching model. I argue that ability to accommodate context effects speaks in favour of the feature-matching model but Tversky's theory of PS is possibly too far-fetched or idealised.

4.2. Feature-matching

In Tversky's (1977) theory, feature-matching is a process of PS. The process is described by a function that takes sets of features as inputs and outputs a linear measure of their overlap. Figure 4.1 illustrates this with schematic faces. a and b

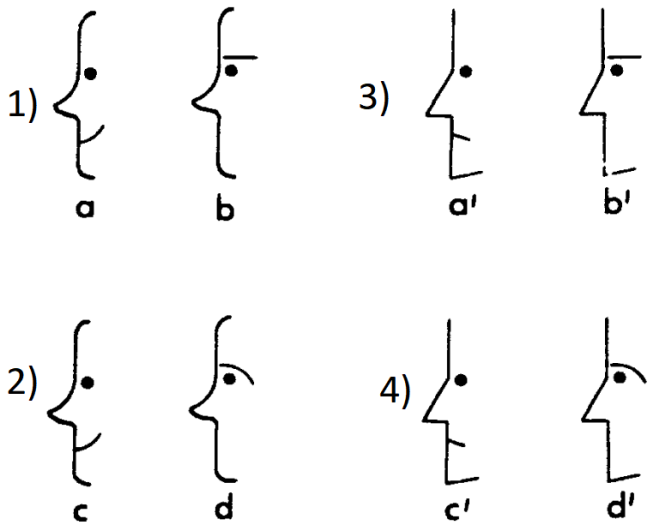


Figure 4.1.: 4 pairs of schematic faces. From "Features of similarity," by Tversky, 1977, Psychological review 84(4), p. 331. Copyright 1977 by the American Psychological Association. Adapted with permission.

overlap with respect to one feature—the round profile. They divide with respect one feature that is distinct to a—the smile—and one feature that is distinct to b—the straight eyebrow. According to Tversky's theory of PS as feature-matching, the PS of two objects is the extent to which they 'match' each other on a linear scale of discrete features. It is apparent that Tversky's theory of PS is mathematically different from Shepard's (1987) theory because the former relies on a set-theoretic description of the data, not on geometric quality dimensions (section 3.3.2).

More formally, the feature-matching function is defined as follows. Let $\Delta = \{a, b, c, ...\}$ be the domain of objects. $A, B, C, ...$ is the set of sets of features, where each feature set is associated with an object from Δ. (For example, A is the set of the features that are associated with the object a.) The feature-matching function characterises PS as a set-theoretic difference between two objects and takes the following general form.

$$s(a, b) = F(A \cap B, A - B, B - A). \tag{4.1}$$

Equation 4.1 says that for a pair of objects, a and b, the PS between these objects is a function, F, of

the intersection of their features, $A \cap B$,

the disunion, $A - B$, of the features that are distinct to a without b, and

the disunion, $B - A$, of the features that are distinct to b without a.

Figure 4.2 illustrates the relations between these sets in the faces example.

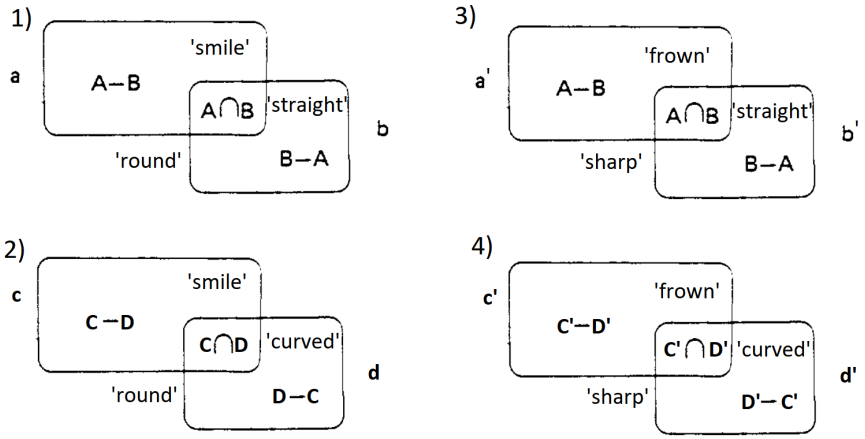


Figure 4.2.: Sets of common and distinct features for each pair of faces from figure 4.1.

4.2.1. Key assumptions

The first assumption of Tversky's model is that sets of common and distinct features are discrete elements. This implies that features must be expressed by integers (i.e., countable whole numbers). For example, the set of possible eyebrows that a face has is countable and this set can be isolated from other sets of facial features[1] A face cannot have 2.5 features.

The second assumption is that the origin of the feature space is a data base of objects, $\Delta = \{a, b, c, ...\}$, where Δ represents the stimuli given in a similarity-judgement experiment. This assumption implies that there exists a process prior to feature-matching, which extracts a set of features associated with each object in the data base. For example, if the data base is $\Delta = \{\text{a, b}\}$, where a and b represent the faces in figure 4.1, the process might extract the features $\langle R, SE, S\rangle$, where R represents the feature *round profile*, SE represents the feature *straight eyebrow* and S represents the feature *smiling*. The extraction process somehow links $\langle R, SE, S\rangle$ to a and b.

On Tversky's approach, when these assumptions are met, the feature-matching algorithm can proceed by counting the number of links that the objects do or do not already share. This illustrates that in contrast to Shepard, Tversky is not concerned with the problem of perception (i.e., how the agent arrives at a stable representation of an object). The goal is to explain how PS judgements are computed once stable object representations (as stored in the data base) are in place, and once the relevant sets of features have been extracted from the objects. In the next section, I motivate this approach and explain how Tversky's model accommodates the directionality associated with judgements of similarity.

[1] A contrasting example is the set of possible colours of a face, which is uncountable because it cannot be isolated: a face has varying shades of colour, which transition gradually into each other. However, this understanding of sets of features leaves open the possibility that a set of features is countable yet cannot be isolated.

4.3. Why feature-matching?

The feature-matching model is motivated by the aim to avoid three shortcomings of the geometric model. Firstly, the geometric model relies on the metric axioms but these, particularly the symmetry axiom, are sometimes violated. Secondly, the geometric model does not explain the directionality associated with judgements of similarity. Thirdly, the geometric model does not capture the full range of possible PS functions. I walk through each shortcoming in turn.

4.3.1. Violations of the metric axioms

The geometric model is ignorant to directionality in the following sense. By definition, the metric distance between two points, a and b, is always the same as the metric distance between b and a. On the basis of the symmetry axiom, the geometric conception lets us expect that subjects' judgements of similarity should be non-directional. That is, we would expect that subjects judge that a is as similar to b as vice versa. Implicit in this expectation is the assumption that directionality is the opposite of symmetry. We can then say that a PS judgement of two objects is symmetric (i.e., non-directional) if and only if it is independent of the order in which the two objects are compared. For instance, a geometric function of the distances between cities on the world map is symmetric (non-directional) because it is ignorant of whether distance is measured from the point representing Tel Aviv to the point representing New York or whether distance is measured from New York to Tel Aviv instead. The function that computes the distance should give the same results in either direction; it should be symmetric. The geometric model illustrates the same ignorance.

Tversky refutes the geometric model on the basis of empirical evidence that seems to show that subjects' judgements of similarity are sometimes directional in this sense. His case in point are observations such as the following. He reports that phenomenon that people are more likely to judge (S1) Tel Aviv to be similar to New York and they are less likely to judge (S2) New York to be similar to Tel Aviv. He also reports that people commonly prefer to say that (S3) 'North Korea is like Red China' instead of saying that (S4) 'Red China is like North Korea.' (Tversky, 2004, p. 8) He extends this analysis to further domains of evidence, showing that similarity statements comparing geometric figures or Morse Code signals are directional as well (Tversky, 1977, pp. 333-334).

Before looking at the details of his explanation of these cases, it should be noted that Tversky uses the term 'directional' in two ways. On the one hand, he uses this term to refer to empirical similarity-judgement data (e.g., the observation that people commonly say that Tel Aviv is more similar to New York than the other way around). In this way, directionality describes the objective behaviour associated with similarity statements (e.g., statements about how similar New York is to Tel Aviv and vice versa). On the other hand, it is implicit in Tversky's explanation of these effects that directionality is also a property of subjects'

mental states about the objects (e.g., how subjects represent the relation between Tel Aviv and New York in their minds). In Tversky's theory, the directionality in the observable similarity statements is structurally the same as the directionality in the internal PS judgements. He seems to think that the latter are verbal expressions of the former. My interpretation is that Tversky understands directionality so that the internal judgements of PS take the form of beliefs that can be understood as propositions in a language of thought (Fodor, 1975, 2008). Nevertheless, conceptually, the two notions should be distinguished. In the following paragraphs, I explicate a possible distinction between them. I start with PS judgements.

Definition 4.3.1 (Directionality, PS judgements)**.** A pair of PS judgements, $S(a, b)$ and $S(b, a)$, is directional whenever $S(a, b) \neq S(b, a)$, where S describes a belief about the relation between a and b.

According to this definition, an example for a PS judgement is the case in which a subject judges the similarity between Tel Aviv and New York to be different from the similarity between New York and Tel Aviv. This judgement is directional; the subject's belief about the relation between Tel Aviv and New York is not equal to the subject's belief about the relation between New York and Tel Aviv.

I contrast this with Tversky's description of directionality in similarity statements. On Tversky's (1977, p. 328) account, a directional similarity statement typically has the (grammatical) form 'a is like b' (1) or 'b is like a' (2). Correspondingly, a non-directional similarity statement should have the form (3) 'a and b are alike.' Recapitulate that it is assumed that there is a structural correspondence between similarity judgements and similarity statements. On this basis, Tversky's illustrations implicitly suggest that the corresponding similarity judgements take the form $S(a, b)$ in (1) and $S(b, a)$ in (2) when $S(a, b) \neq S(b, a)$. The form of (3) implies that $S(a, b) = S(b, a)$, so that the corresponding similarity judgement is symmetric. I define the directionality in similarity statements as follows.

Definition 4.3.2 (Directionality, similarity statements)**.** A similarity statement, 'a is like b', is directional if, when asserted, its counterpart, 'b is like a', is simultaneously denied. Formally: $s(a, b) \neq s(b, a)$.

In light of definition 4.3.2, the explanation of directionality builds on changes in the meaning of similarity statements. For example, people judge the similarity between Tel Aviv and New York to be different from the similarity between New York and Tel Aviv because the terms referring to these objects in people's expressions of similarity statements take on different semantic roles and this creates a difference in the meanings of the statements "Tel Aviv is like New York" (S1) and "New York is like Tel Aviv" (S2). In S1, Tel Aviv has the role of the subject and in S2, Tel Aviv plays the object, while New York takes on the role of the object in S1 and the role of the subject in S2. Tversky

seems to understand the change in the meaning from S1 to S2 to be somehow caused by a difference in the content associated with people's internal similarity judgements. This seems somewhat plausible under the implicit assumption that similarity statements are verbal expressions of similarity judgements (e.g., beliefs with the same structural form). Thus, in the actual example, the explanation of the current example is that the similarity statement about Tel Aviv is directional (i.e., s(Tel Aviv, New York) $\neq$ s(New York, Tel Aviv)) because subjects' inner similarity judgement is directional (i.e., S(Tel Aviv, New York) $\neq$ S(New York, Tel Aviv)). More generally, when a pair of similarity statements is directional, this observation is the expression of the directionality between a pair of similarity judgements. Correspondingly, effects showing that pairs of similarity statements are sometimes directional indicate that pairs of judgements of similarity are sometimes directional.

This outline of Tversky's understanding of directionality makes his claim that directionality violates the assumption that PS relies on the symmetry axiom more transparent. Provided that directionality is the opposite of symmetry and that there is a structural correspondence between similarity statements and similarity judgements, the hypothesis that pairs of statements of similarity are sometimes directional implies that judgements of similarity are sometimes asymmetric. Therefore, Tversky's finding of evidence for this hypothesis is evidence against the universal truth of the assumption that PS is a function of geometric distance.

Note that Tversky attempts to refute minimality and triangle inequality as well. He attempts to refute minimality based on two claims. (1) Self-similarity is not always the same. (2) Self-similarity is typically less than 1. Both claims are confirmed by the averaged data in the diagonal entries in Rothkopf's similarity matrix (section 3.4, figure 3.2, left).[2] However, these results do not violate what is implied by the minimality axiom. Firstly, it does not follow from minimality that self-similarity must be the same *across* subjects and stimulus pairs. Secondly, the most obvious prediction of the minimality axiom can be confirmed; the diagonal entries in Rothkopf's similarity matrix (figure 3.2, left) exceed the off-diagonal entries.

Tversky endeavours to refute triangle inequality on the basis of a simple counterexample, which centers on the similarities among Cuba, Russia and Jamaica. Recapitulate that triangle inequality says that for any triadic relationship, the individual distances between any pair of points must be smaller than or equal to the sum of the other two alternative distances. Correspondingly, the judged dissimilarity between Jamaica and Russia must be smaller than the joint dissimilarities between Cuba and Jamaica and Cuba and Russia. Formally, if triangle inequality holds: δ(Jamaica, Russia) $\leq$ δ(Jamaica, Cuba) + δ(Cuba, Russia).

[2]The diagonal of this matrix indicates similarity judgements associated with objectively identical stimuli, $(a, a), (b, b), (c, c)$, etc. For example, if b is the signal '-...', then the index (b, b) represents the typical similarity judgement for the identical signal pair '-...' and '-...' . In figure 3.2 (left), these indexes vary from 83 to 96 percent. Therefore, the averaged identification probability is not constant across identical pairs.

Tversky's argument against triangle inequality rests on the premise that triangle inequality implies transitivity[3] and the hypothesis that PS cannot be expected to be always transitive. In the example, Jamaica is geographically similar to Cuba and Cuba is politically similar to Russia but Jamaica and Russia are not similar at all (or, to put it differently, Jamaica and Russia are extremely dissimilar). If PS was transitive, Russia should be in some respect similar to Jamaica. Tversky drives this example a bit further, suggesting that $\delta(\text{Jamaica, Russia}) > \delta(\text{Jamaica, Cuba}) + \delta(\text{Cuba, Russia})$, which is roughly the opposite of what triangle inequality predicts. However, the caveat of Tversky's example is that it is a thought experiment and still needs empirical validation.[4]

Considering these arguments, I think that Tversky's case against the geometric model is strongest with respect to the directionality associated with pairs of similarity statements and the hypothesis that directionality is a property associated with the internal structure of pairs of PS judgements.

Taken together, this section has explained Tversky's (1977) claim that the observation of directionality in similarity statements is evidence against the axiom of symmetry. I have explicated Tversky's implicit assumptions about the relationship between directionality and symmetry and the systematic correspondence between similarity judgements and their expression in similarity statements. On this basis, I have outlined how Tversky's theory of PS as feature-matching possibly explains the apparent violations of the symmetry axiom. In conclusion, if Tversky's assumptions are correct, then Shepard's (1987) geometric model of PS cannot represent all aspects of PS. The next section outlines Tversky's more precise explanation of directionality effects.

4.3.2. Origins of directionality

Tversky's theory of PS explains directionality under the assumption that three conditions are simultaneously met.

1. The similarity-judgement task is formulated in a directional way. The task, 'how similar are a and b to each other?' is non-directional but 'how similar is a to b?' is directional.

2. One of the distinct sets of features associated with the objects is more salient than the other. In the model, this requires that $f(A - B) \neq f(B - A)$.

[3] If a is quite similar to b and b is quite similar to c, then a cannot be very dissimilar from c either. All other things being equal, the further c departs from a, the more likely it becomes that the triangle axiom will be violated.

[4] Similarly, Müller-Trede, Sher, and McKenzie (2015, p. 280) argue that "no clear triangle inequality violations have been empirically demonstrated to date" Yearsley, Barque-Duran, Scerrati, Hampton, and Pothos (2017, p. 27). add to this observation in pointing out the difficulty that "there is currently no precise notion of how the triangle inequality translates into a constraint for similarities, as opposed to dissimilarities."

3. The focusing hypothesis: if conditions 1 & 2 hold, then the sets of distinct features should be weighted unequally in the feature-matching process, so that $\alpha \neq \beta$.

Following Tversky (Tversky, 1977, pp. 331-332), the feature-matching model can explain directionality whenever these three conditions are met. I outline the intuitive explanation of this here and illustrate the details of how this works in section 4.3.3. Roughly, the intuitive explanation is that the effect of directionality occurs because subjects change their focus of attention to the objects' distinct features when the experimental task changes in design. In particular, when conditions 1-3 are met, the objects become relatively more or less distinct with a change in their relative order in the comparison. One interpretation of this dynamic is offered in Tversky and Gati (1978, pp. 81, 85), who argue that the objects under comparison become relatively more or less distinct because their semantic roles in the similarity statements change. On this understanding, subjects' shift of attention is elicited by their interpretation of the semantic roles of the referring terms in similarity questions. The assumption is that people pay generally more attention to whatever stimulus in the pair takes the place of the object (e.g., New York in S1), and less attention to the stimulus that plays the role of the subject (e.g., New York in S2). When the more salient stimulus (e.g., New York) takes the role of the object, the distinctiveness (dissimilarity) between the objects is strengthened (e.g., in S1). When the more salient stimulus takes the role of the subject, the objects are relatively less distinct (e.g., in S2). More generally, a change in the relative position of two objects of different salience (as indicated by differences in the cardinalities of their distinct sets of features) in the comparison produces the directionality effect.

To illustrate Tversky's explanation of directionality, I forestall aspects of the contrast model (section 4.3.3). In the model, $S(a, b)$ is the linear difference between the common and distinct sets of features associated with a and b. When $\alpha > \beta$, then in S1, the set of features distinct to Tel Aviv obtain more weight than the set of features distinct to New York. In S2, the set of features distinct to New York obtain more weight than the set of features distinct to Tel Aviv. Since Tel Aviv is, intuitively, associated with fewer distinct features (it is relatively less salient), its position in second place in the contrast model (corresponding to S2) subtracts a smaller amount of dissimilarity from the set of common features. Table 4.1 summarises these results.

Taking stock, this section has explained directionality with the feature-matching function, F (equation 4.1) and with Tversky's focusing hypothesis ($\alpha \neq \beta$). Recapitulate that in equation 4.1, F takes sets of features to a similarity measure but it does not specify the mathematical operation that combines these sets (e.g., addition, subtraction, multiplication or division). The next section explains two variants of the feature-matching model—the contrast model and the ratio model.

New York (salient)	Tel Aviv (not salient)	PS
object	subject	high
subject	object	low

Table 4.1.: Illustration of the influence of relative position between objects of different salience in feature-matching. Relative position is indicated as semantic role (subject or object) in a similarity statement. The prediction of amount of PS (high vs low) is based on the additional assumptions that the task is directional and the focusing hypothesis.

4.3.3. Diversity of directionality

Tversky (1977, p. 322) describes the contrast model with the following mathematical function.

$$S(a,b) = \theta f(A \cap B) - \alpha f(A - B) - \beta f(B - A), \text{ for some } \theta, \alpha, \beta \geq 0. \quad (4.2)$$

The model represents $S(a,b)$ as a linear combination of the shared and distinct features associated with the object pair a and b. The parameter f is a non-negative scale over a given set-theoretic space, which relates the function S to the features of the objects. f takes the set of common or distinct features as an input and outputs a value for the cardinality of that set. The weights, θ, α and β are positive constants between 0 and 1. They determine how much each set of features contributes to the overall measure of PS associated with a and b. If $\theta = 1, \alpha = \beta = 0$ then similarity depends only on the set of common features. Conversely, if $\theta = 0, \alpha = \beta = 1$ then similarity depends only on the sets of distinct features.

In the contrast model, a change in the PS associated with a and b is a result of changes in the absolute cardinality of sets of features. When keeping the cardinality of common features constant, directionality effects can be accommodated by either swapping the order of the objects and their associated distinct features or by changing the weights associated with the corresponding sets of distinct features. In other words, under the focusing hypothesis (i.e., $\alpha \neq \beta$), an effect of directionality is either produced by a change in the relative difference between the weights (e.g., by changing $\alpha > \beta$ to $\alpha < \beta$), or by a change in the relative position of the sets of distinct features (i.e., by swapping the positions of the terms $f(A - B)$ and $f(B - A)$). Table 4.2 illustrates these effects.

Table 4.2 shows that, *under the focusing hypothesis ($\alpha \neq \beta$)*, a change in the relative weight between a pair of two sets of distinct features is enough to produce directionality. This can be seen when comparing row 1, where $\alpha < \beta$, and row 5, where $\alpha > \beta$, while keeping all other parameters fixed. A comparison between

	$f(A \cap B)$	$f(A - B)$	$f(B - A)$	α	β	$S(\cdot, \cdot)$
1.	100	60	10	2	5	-30
2.	100	10	60	5	2	-70
3.	100	60	10	2	2	-40
4.	100	10	60	2	2	-40
5.	100	60	10	5	2	-180
6.	100	10	60	2	5	-220

Table 4.2.: Illustration of the accommodation of directionality ($S(a, b) \neq S(b, a)$) with different parameter settings in the contrast model.

rows 1 and 6 illustrates that, under the focusing hypothesis, only a change in the relative position of the distinct sets of features suffices likewise to evoke an effect of directionality. However, changing the relative order of the objects in the comparison (corresponding to a change in the relative position of the sets of distinct features in the model) will not suffice to evoke directionality if the focusing hypothesis is denied (i.e., when $\alpha = \beta$). This is illustrated in the comparison between rows 3 and 4. Thus, the focusing hypothesis needs to be accepted to produce effects of directionality with the contrast model.

The second variant of the feature-matching model is the ratio model. This model represents $S(a, b)$ as a ratio of the number of the common features to the total number of the common and the distinct features associated with an object pair. Tversky describes this model with the following mathematical function.

$$S(a, b) = \frac{f(A \cap B)}{f(A \cap B) + \alpha f(A - B) + \beta f(B - A)}, \text{ for some } \alpha, \beta \geq 0. \qquad (4.3)$$

In equation 4.3, PS is normalised and S takes on values between 0 and 1. This implies that the sum of the values in the denominator is fixed relative to the measure of similarity between a and b.

The ratio model can accommodate directionality effects as well. For comparisons with the same pair of objects, PS increases when there are more common than distinct features and when one set of distinct features is more important than the other (i.e., $\alpha \neq \beta$). PS decreases when the opposite is the case. Table 4.3 illustrates this with numbers. Under the focusing hypothesis (i.e., $\alpha \neq \beta$), only changing the relative weights, while all other parameters are fixed, will evoke directionality in the ratio model. This is illustrated in the comparison between rows 1 and 3. A comparison between rows 1 and 6 illustrates that a change in the relative position of the objects and their associated distinct sets of features is,

ceteris paribus, sufficient to generate an effect of directionality. This result holds only when the focusing hypothesis ($\alpha \neq \beta$) can be accepted. This is illustrated in the comparison between rows 3 and 4 (table 4.3).

	$f(A \cap B)$	$f(A - B)$	$f(B - A)$	α	β	$S(\cdot, \cdot)$
1.	100	60	10	2	5	.37
2.	100	10	60	5	2	.37
3.	100	60	10	2	2	.42
4.	100	10	60	2	2	.42
5.	100	60	10	5	2	.24
6.	100	10	60	2	5	.24

Table 4.3.: Illustration of how directionality ($S(a, b) \neq S(b, a)$) can be accommodated by setting parameters differently in the ratio model.

Taken together, equations 4.2 and 4.3 look quite different. How can they be understood as instances of a common feature-matching process? In some sense, both models belong to an overarching description of what it means for two objects to be similar. Firstly, in both models, S ('similarity') is an interval scale over a pair of objects, a and b, and preserves the similarity order between a and b. Whenever it is true that $s(a, b) \geq s(c, d)$, then it will be true that $S(a, b) \geq S(c, d)$. For example, when the observed similarity between an apple and a pear is greater than the similarity between a banana and a peach, the scale S maintains this information, regardless of the particular numbers that it assigns to the similarities of these objects. Secondly, in both models, this relationship is linear. S increases linearly with a linear increase of the number of common features of a and b and S decreases linearly with an increase in the number of distinct features. Finally, both models can explain directionality only under the additional assumption that conditions 1-3 in section 4.3.2 are met. Most importantly, both models rely on the truth of the focusing hypothesis ($\alpha \neq \beta$), as is illustrated in tables 4.2 and 4.3.

In another sense, the contrast and ratio models are differently informative about what it means for a and b to be relatively (dis-)similar. In equation 4.3, a change in the relative position of the objects (e.g., row 1 vs row 6, table 4.3) has the same effect as a change in relative weight (e.g., row 1 vs row 5, table 4.3). In contrast, in equation 4.2, a change in relative position (row 1 vs row 6 in table 4.2) has a different effect as a change in relative weight (row 1 vs row 5 in table 4.2). The reason for this is that in the ratio model (equation 4.3), a change in S is relative to the number of all features considered, while in the contrast model (equation 4.2), the change is a function of the absolute differences between the numbers attached to common and distinct features. In other words, according to the contrast model, PS is an absolute relationship between a and b, while

according to the ratio model, PS indicates how much more or less similar a and b are in relation to the associated overall cardinality of features.

Taken together, I have explained that Tversky's accommodation of directionality with either the contrast or the ratio models relies on the three conditions that the task is directional, that the objects are differently salient and that the subject pays more attention to one object than to the other (i.e., $\alpha \neq \beta$). I have illustrated this with the cities and countries examples. In these examples, Tversky's explanation of directionality relies on an interpretation of the semantic roles that are played by the objects. However, it is possible that shifts in attention may be elicited in a more general way. From the perspective of the feature-matching model, the focusing hypothesis should be true when conditions 1 & 2 are met, and these conditions are independent of an interpretation of the objects' semantic roles. In the next section, I discuss one of Tversky's tests of the generality of the focusing hypothesis under conditions 1 and 2. This is Tversky's analysis of Rothkopf's (1957) Morse code data (section 4.4). Following this analysis, directionality is relatively independent of the type of stimuli. I illustrate that under the assumption that shorter signals are less salient than longer signals, and when $\alpha > \beta$ is fixed, PS in the model will be higher when the longer signal takes the second position while the shorter signal takes the first position in the comparison (and PS will be lower when this order is reversed, under the assumption that $\alpha > \beta$).

4.4. A case study: directionality in similarity judgements of Morse Code signals

Tversky's analysis of the Morse Code data (Tversky, 1977, p. 336) serves to test the focusing hypothesis. To recapitulate, the hypothesis says that the distinct sets of features of a pair of objects in a similarity-judgement task are differently important to a subject, depending on the relative position of the objects in the comparison. In the feature-matching model, this translates to the assumption that $\alpha \neq \beta$. To test the focusing hypothesis with an analysis of Rothkopf's (1957) data, Tversky concentrates on only those pairs of Morse Code signals, i and j, that are judged to be differently similar. When analysing these pairs, Tversky considers at two factors. The first factor is the order of a signal within a pair (i.e., the relative positions of i and j). The second factor is the temporal length of the signals (i.e., how many components each signal has). He then addresses two questions:

1. Is a subject more (or less) likely to confuse i and j if i is presented first and j is presented second (respectively, vice versa)?

2. Is a subject more (or less) likely to confuse i and j if i is temporally longer than j (or vice versa)?

The first question serves to tests the hypothesis that similarity judgements are directional. The second question serves to test the hypothesis that the directionality of a subject's similarity judgement is guided by different attentional weights on sets of distinct features. When accommodating directionality in the Morse Code data, Tversky uses the contrast model. I have illustrated in section 4.3.2 that the corresponding explanation of directionality effects relies on three additional assumptions. In Tversky's analysis of the Morse Code data, these are the following assumptions.

(a) Temporally longer signals (denoted by 'p') are more salient than temporally shorter ones (denoted by 'q').

(b) The focusing hypothesis (i.e., $\alpha \neq \beta$): differences in relative salience induce differences in subjects' foci of attention to the distinctive aspects of each signal.

(c) The task is directional: when q is presented prior to p, this is equivalent to q being the subject and p being the object in the comparison.

Tversky's analysis shows that, when (a)-(c) are met, the contrast model can be used to accommodate effects of directionality in the Morse Code data. In particular, the model can be used to accommodate the observation that the probability to confuse the signals '..' (q) and '.–.' (p) is sometimes greater than the probability to confuse the signals '.–.' and '..'. From the perspective of the theory of PS as feature matching, a possible explanation of this effect would be that if conditions (a)-(c) are met, then the distinct features associated with p decrease the PS of the pair relatively more than the distinct features associated with q in the first case than in the second case. Two aspects of Rothkopf's (1957) data seem to support Tversky's assumption that conditions (a)-(c) are met and that the contrast model can be used to accommodate these effects. Firstly, the probability to confuse a signal pair was usually higher when the shorter signal appeared prior to the longer signal. This seems to be relevant to the first and second conditions: depending on the relative length of the signals, the objects become differently salient. Secondly, the order in which the signals were presented made a difference to how likely is was to confuse them. This seems to relate to the third condition, that the order of the signals is important to the difference in their salience.

Kind of relationships	Number of cases	Percentage of total no. of cases
$s(q,p) > s(p,q)$	336	0.61
$s(p,q) > s(q,p)$	181	0.33
$s(q,p) = s(p,q)$	38	0.07

Table 4.4.: Overview of the effects of directionality in Rothkopf's Morse code data, based on Tversky's (1977, p. 336) analysis.

A summary of these findings is presented in table 4.4. Row 1 indicates that relative to all 555 analysed cases, it was more likely to confuse q and p if q is

presented prior to p. Row 2 indicates the opposite, that it was more probable to confuse q and p if p is presented prior to q. $s(q,p)$ exceeds $s(p,q)$ in 61 percent of all trials[5].

These results support Tversky's assumptions. Firstly, the relative differences between the results in the third columns associated with rows 1 and 2 confirms the assumption that the signals' relative temporal length affects their relative salience. Secondly, the results in row 1 confirm the assumption that the signals' relative position affects the directionality in their comparison. Overall, table 4.4 illustrates that the effect of directionality is a pattern in the data. There are 517 out of 555 pairs of similarity judgements that are directional and only 38 out of 555 cases in which confusion probability is symmetric despite a change in the relative order of p and q. Intuitively, this pattern needs explanation.

In the following paragraphs, I use Tversky's contrast model to elucidate why the directionality effect might occur. For any pair of signals, (p,q), let P and Q stand for their associated features, respectively. $(P \cap Q)$ is the set of features common to both p and q, $(P - Q)$ is the set of features distinct to p and $(Q - P)$ is the set of features distinct to q. Following the contrast model (equation 4.2), the similarity between p and q can be computed as follows.

$$\text{A. } S(p,q) = \theta f(P \cap Q) - \alpha f(P - Q) - \beta f(Q - P).$$

And the similarity between q and p can be computed as follows.

$$\text{B. } S(q,p) = \theta f(Q \cap P) - \alpha f(Q - P) - \beta f(P - Q).$$

For example, let p stand for the signal '..–' and let q stand for the signal '..'. Assuming that there is no focusing, so that $\alpha = \beta$, the contrast model can be used to predict that $S(p,q) = S(q,p)$.

$$\text{C. } S(p,q) = \theta f(P \cap Q) - \alpha f \times (1) - \beta f \times (0) = \theta f(P \cap Q) - 1.$$
$$\text{D. } S(p,q) = \theta f(Q \cap P) - \alpha f \times (0) - \beta f \times (1) = \theta f(P \cap Q) - 1.$$

In both C and D, PS is a function of the set of shared features minus one set of distinct features. Thus, without any additional assumptions, the PS between p and q is symmetric in C and D. This changes upon introducing the focusing hypothesis (i.e., $\alpha \neq \beta$), under which the contrast model accommodates differences in the ordering effects. Assuming that $\alpha = 3$ and $\beta = 1$, the contrast model can be used to predict that $S(q,p) > S(p,q)$, such as in the following case.

[5] These numbers add up to more than 100 because they are rounded. When using the exact numbers from Rothkopf's data, the percentages in the third column add up to 100.

$$\text{E. } S(p,q) = \theta f(P \cap Q) - 3 \times f \times (1) - 1 \times f \times (0) = \theta f(P \cap Q) - 3.$$
$$\text{F. } S(p,q) = \theta f(Q \cap P) - 3 \times f \times (0) - 1 \times f \times (1) = \theta f(P \cap Q) - 1.$$

In E and F, PS between p and q is directional: in E, 3 sets of features are subtracted from the overall set of common features, while in F, only 1 set of features is subtracted. The features distinct to p (i.e., temporal length) contribute more to the dissimilarity because the features distinct to p weigh heavier in the PS they subtract from the set of common features.

Interim conclusion

The discussion so far has established the claim that the feature-matching model accommodates directionality effects. These effects are at odds with the symmetry assumption of Shepard's (1987) geometric model of PS (section 3.3.1). A similarity judgement associated with two objects, a and b will sometimes be directional so that the PS between a and b is different from the PS between b and a.

Following Tversky's model, I have given two possible explanations for this phenomenon (section 3.3.2). The first explanation is that directionality is determined by the relative salience of the objects and effects of directionality can be expected if a is more salient than b or vice versa. This explanation alludes to subjects' background knowledge about the objects' relative salience and their semantic roles in similarity statements. If a plays the role of the object and b plays the role of the subject but b is more salient than a, it can be expected that a and b will be less psychologically similar than if a was the object and b the subject in the comparison. On Tversky's account, these semantic roles are determined by factors implicit in how the experimental task is formulated. The second explanation is that directionality depends on the relative number of distinct features associated with each object so that the PS between a and b will be smaller than the PS between b and a if a has more distinct features than b. This explanation alludes to the attentional mechanisms involved when comparing a pair of objects. In other words, an object with a larger number of independent features will contribute more to making the pair dissimilar. How much an object contributes to the dissimilarity depends on the subjects' sensitivity to the object's distinct features.

In section 3.3.3, I have shown in detail how the feature-matching theory of PS accommodates directionality effects with the contrast and ratio models. Section 3.4, has illustrated Tversky's explanation of directionality effects in Rothkopf's (1957) Morse Code data. The next section connects Tversky's theory of PS to the phenomenon of PC. I focus on an explanation of how the feature-matching model in conjunction with Tversky's diagnosticity principle (s.b.) can accommodate effects of context on PC behaviour.

4.5. Feature-matching and categorisation

Tversky's diagnosticity principle asserts that for a given set of objects that can be grouped into different categories, a replacement (addition or deletion) of category members can alter how the remaining objects in the array will be categorised. The basic assumption of the principle is that objects are grouped with respect to the diagnostic value of their associated features. A feature has diagnostic value for a grouping if and only if it is significant for the group. For example, the diagnostic value of *being real* is low for the category *animal* in a context in which there are only real animals because *being real* is insignificant for any grouping amongst the real animals; all available animals in that context are real. In contrast, in an expanded context that includes also legendary animals like centaurs, mermaids or a phoenixes, the diagnostic value of the feature *being real* increases. In this expanded context, *being real* becomes diagnostic of the category *animal* (cf. Tversky, 1977, p. 342). I illustrate Tversky's diagnosticity principle with two examples (figures 4.3 and 4.4).

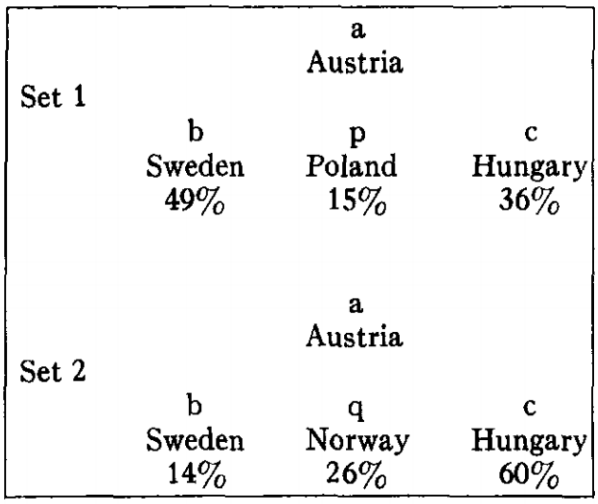


Figure 4.3.: Illustration of the diagnosticity principle with two sets of four countries. The percentage of subjects who grouped any test country together with Austria (the target) is indicated below that country. The average groupings were: [a,b] & [p,c] in set 1 and [a,c] & [b,q] in set 2. From "Features of similarity," by Tversky, 1977, Psychological review, 84 (4), p.343. Copyright 1977 by the American Psychological Association. Reprinted with permission.

The first example is an experiment from Tversky & Gati (1978, pp. 92-95) on the comparison between sets of countries. In one condition (1), Austria (a) was compared to Sweden (b), Poland (p) and Hungary (c). In another condition (1), Austria (a) was compared to Sweden (b), Norway (q) and Hungary (c). Both conditions offer different contexts: (1) offers a political interpretation of the similarities between these 4 countries and (2) offers a classification based on their geographic similarity. This difference is induced by the replacement of

Poland (p) in (1) with Norway (q) in (2). Subjects had to group the set of 4 countries into pairs. The results of Tversky & Gati's experiment are indicated in the percentages in figure 4.3, which is taken from Tversky's (1977, pp. 342-343) report of this experiment. These results support the diagnosticity principle. In condition (1), subjects were most likely to categorise Sweden and Austria into one category and Poland and Hungary into another. In condition (2), subjects were most likely to classify Austria and Hungary into one category and Sweden and Norway into another. Tversky & Gati (1978, pp. 92-94) interpret these results to indicate that the categorisation of the countries into clusters changes with the context because the context changes the diagnostic values associated with features of these countries[6]. In (1), political features are diagnostic of the categorisation of Austria with Sweden and Poland with Hungary. In (2), geographic features are diagnostic of the categorisation of Austria with Hungary and Sweden with Norway.

Note that the stimuli in Tversky & Gati's experiment are not purely perceptual; they are names of countries and they clearly require the subject to know the geographic and political situation of the countries to categorise them. This knowledge is not purely perceptual—subjects do not perceive the geographic and political situation of the countries in the experiment. Thus, one may wonder whether this example clearly illustrates the connection between Tversky's diagnosticity principle and the phenomenon of *perceptual* categorisation.

The second example is an experiment reported in Tversky (1977, p. 342) on the comparison between sets of schematic faces (figure 4.4). This example is supposed to illustrate more clearly that Tversky's diagnosticity principle is relevant to PC. In one condition (1), the objectively neutral target face (a) is compared with faces b, p and c. In another condition (2), a is compared with b, q and c. Intuitively, (1) is a 'smiling' context while (2) is a 'frowning' context. As in the previous example, a change in the context is induced by the replacement of the stimuli p and q and a subjects' task was to group the 4 objects into pairs. The results, indicated in the percentages in figure 4.4, support the diagnosticity principle. In (1), subjects were most likely to categorise a and b versus p and c. In (2), subjects were most likely to categorise a and c versus b and q. Tversky's (ibid.) interpretation of these results is that the context as induced by replacement of p and q determines the grouping of non-smiling versus smiling faces in (1) and frowning versus nonfrowning faces in (2). *Smiling* has a greater diagnostic value than *frowning* in (1) than in (2) and the diagnosticity of *frowning* versus *smiling* is greater in (2) than in (1).

The examples illustrate that Tversky's diagnosticity principle is about the context-sensitivity of categorisation. Tversky's (1977, p. 342) rationale for the principle is that "[c]lusters are typically selected so as to maximize the similarity of objects within a cluster and the dissimilarity of objects from different clusters." On this basis, Tversky generates the diagnosticity hypothesis, which is a hypothesis

[6]In this experiment, the context seems to be constituted by the objects themselves (e.g., which countries are compared), it may be explicitly chosen by the experimenter (e.g., when the experimenter asks for a judgement of political versus geographical similarity).

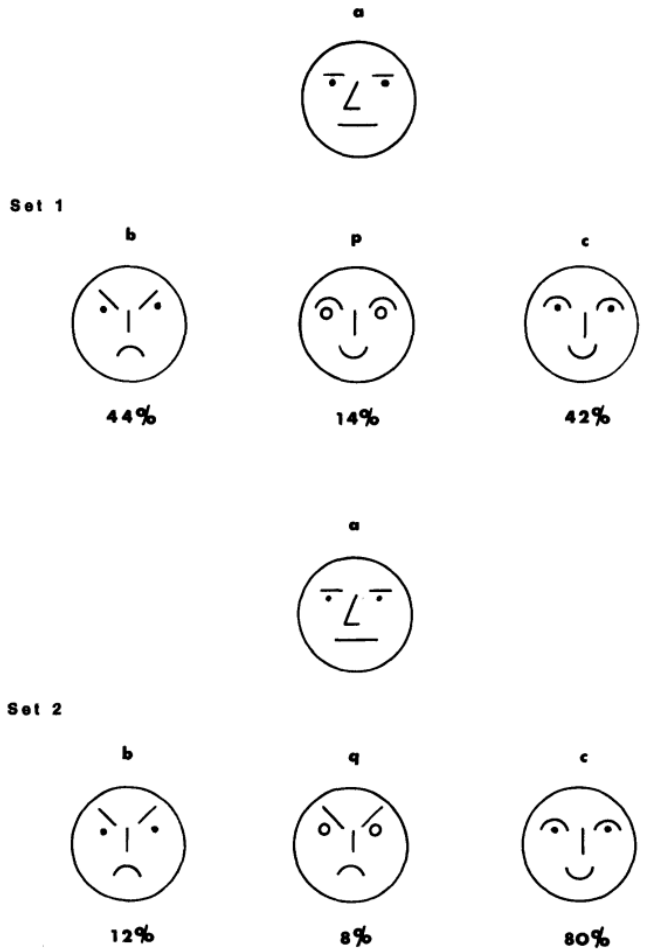


Figure 4.4.: Illustration of the diagnosticity principle with two sets of four schematic faces. The percentage of subjects who selected any face together with the target (a) is presented below that face. From "Features of similarity," by Tversky, 1977, Psychological review, 84(4), p. 341. Copyright 1977 by the American Psychological Association. Reprinted with permission.

about the context-sensitivity of similarity judgements. The diagnosticity hypothesis says that "[a] change of clusters [...] is expected to increase the diagnostic value of features on which the new clusters are based, and therefore, the similarity of objects that share these features" (Tversky, 1977, p. 342). What this means is that features with a higher diagnostic value obtain more weight in the feature-matching model than features with a lower diagnostic value. This can lead to a change in the overall cardinality of the sets of common and distinct features in the comparison. My earlier illustrations of the contrast and ratio models have shown that a changes in the weights and cardinality of sets of features are likely to lead to different outcomes in the feature-matching process. On this basis, it seems plausible to think that a change in the diagnostic value of a set of features associated with two objects in a comparison is likely to induce a change in the similarity judgement associated with the objects. A change in the

diagnostic value of a feature influences the way in which objects are categorised, since diagnosticity depends on the context (e.g., what objects are available in a similarity-judgement or categorisation task). Therefore, Tversky's experiments illustrate that the context indirectly influences categorisation.

If we accept that the diagnosticity hypothesis is true in the above examples, we would predict that $S(a, b)$ in (1) $> S(a, b)$ in (2), and $S(a, c)$ in (2) $> S(a, c)$ in (1). Intuitively, under the assumption that diagnosticity is important for categorisation, we would expect that subjects are more likely to group a with b in (1) but not in (2) and that they are more likely to group a with c in (2) but not in (1). Under the assumptions that subjects' categorisations reflect their internal similarity judgements and they maximise PS among objects within a category and minimise PS among objects across categories, these predictions are confirmed by the results in both examples. In the countries example, in (1), Sweden and Austria are similar in that they are constitutional states while Poland and Hungary are similar in that they are both former communist states. In (2), Austria is more similar to Hungary than to Sweden or Norway; Austria is geographically closer to Hungary in comparison to Sweden and Norway, which are geographically close. In the faces example, people pair a and b more often than a and c in (1) while in (2), a and c are more often paired than a and b.

Taken together, this section has explained Tversky's diagnosticity principle. The basic assumption of the principle is that objects are grouped with respect to the diagnostic value of their associated features. On the basis of the principle, Tverksy had shown that judgements of similarity are context-sensitive. I have discussed two empirical examples from Tversky (1977) and Tversky and Gati (1978) that connect this result to the phenomenon of PC. From the perspective of Tversky's diagnosticity hypothesis, PC can be expected to change with a change in the diagnostic value of a set of features that is associated with a set of perceptual objects when one of the objects is replaced with another from one context to the other. The next section evaluates Tversky's (1977) feature-matching approach to similarity and categorisation.

4.6. An evaluation of the feature-matching approach

My evaluation of Tversky's approach has a negative and a positive side. On the negative side, I argue that Tversky's approach is too idealised because the assumption that features are discrete sets is too simple to accurately describe aspects of similarity and categorisation in the domain of perception. On the positive side, I argue that Tversky's approach is valuable because it offers an explanation of the plausible assumption that patterns of PS and PC often vary with regards to the context. I start with the positive side.

4.6.1. Context-sensitive

The examples in section 4.5 have shown that PS is context-sensitive with regards to explanations of PC. It may be surprising that this is a positive result. In the introduction of chapter 3, I have mentioned Goodman's (1972) criticism that PS is not explanatory because it would vary too much with the context. In that chapter, I have motivated ULG as a case in point against Goodman's criticism. ULG shows that with respect to generalisation, PS is invariant and useful to predict a wide variety of empirical data. But Goodman is also right in his observation that similarity judgements are obviously context-sensitive. Goodman gives an intuitive example:

> [C]omparative judgments of similarity often require not merely selection of relevant properties but a weighting of their relative importance, and variation in both relevance and importance can be rapid and enormous. Consider baggage at an airport checking station. The spectator may notice shape, size, color, material, and even make of luggage; the pilot is more concerned with weight, and the passenger with destination and ownership. Which pieces are more alike than others depends not only upon what properties they share, but upon who makes the comparison, and when [...] Circumstances alter similarities. (Goodman, 1972, p. 445)

The respects in which the spectator, pilot and passenger may judge pieces of baggage to be similar may obviously vary depending on what aspects of the baggage they pay attention to. An adequate model of PS should be able to account for this intuition. Therefore, such a model should represent the aspect that PS varies with the context. Recapitulate that the geometric model of PS does not do this. The geometric representation of distances does not vary with respect to the context because the dimensions are fixed by the modeller. On the contrary, Tversky's contrast model accounts for the observation that similarity judgements change with the context (section 4.5). The following paragraphs support the claim that this is a good outcome for Tversky's theory of PS.

One argument in favour of Tversky's diagnosticity principle is that it offers a possible explanation of why patterns of PC might vary depending on the context. An intuitive explanation is the following. In a political context, Cuba and Russia are similar but in a geographical context, Cuba and Jamaica are similar. Their membership to the same political or geographical category influences the judgement of similarity. Likewise, in a political context, Austria and Sweden are more similar than Austria and Hungary while in a geographic context, Austria is more similar to Hungary than to Sweden. These are two examples for how the feature-matching model in conjunction with the diagnosticity principle explains context effects.

In conjunction with the diagnosticity principle, the feature-matching theory of PS can accommodate changes in the context by adjusting the weights in the contrast model. For example, consider figure 4.1 and assume that people like

smiling faces more than frowning ones, and that they prefer round ones to sharp ones but are indifferent towards the shape of eyebrows. Correspondingly, in 1) and 2), $\theta, \alpha > \beta$ and in 3) and 4), $\theta, \alpha < \beta$, so that, according to the contrast model: $S(a,b) = S(c,d) = 2+2+0 = 4 > S(a',b') = S(c',d') = 0+0+2 = 2$. This accommodates directionality effects in PC. Depending on the distribution of preferences (i.e., the settings of the weights), two different groupings of the faces according to their pair-wise similarities arise. For example, $g_1 = (a,b,c,d)$ and $g_2 = (a',b',c',d')$, or a single group that subsumes all these types of faces. If the preferences (e.g., the faces' saliency and order) change, the similarity-orders amongst the pairs change, and consequently, it is reasonable to change the groupings. For example, if people like smiling faces more than frowning ones, and they prefer round eyebrows to straight ones but they are indifferent towards a face's profile, then $S(a',b') = 1 < S(a,b) = S(c',d') = 3 < S(c,d) = 5$. Thus, intuitively, only the two pairs in the middle shall be grouped according to these similarity judgements.[7]

Nevertheless, it should be noted that only the contrast model but not the ratio model is sensitive to changes in the context. The ratio model is not context-sensitive because different weights do not produce differences in S. To illustrate this, consider the case in which people have a preference for smiling and round faces over frowning and sharp faces but are indifferent towards eyebrow shape and assume that $\alpha = 2$ & $\beta = 0$ in $S(a,b)$ & $S(c,d)$ and that $\alpha = 0$ & $\beta = 2$ in $S(a',b')$ & $S(c',d')$. These differences lead to an ordering of the Ss in the contrast model but the ratio model stays indifferent towards this order so that $S(a,b) = S(c,d) = S(a',b') = S(c',d') = 1/3$.[8] This order stays constant even if various different values for α and β are considered (table 4.3). The only way in which the ratio model could accommodate differences in S is by changing the absolute sizes of the sets of the common and the distinct features for each object pair. Yet, these are fixed for each pair and do not vary with the context. Thus, contrary to the contrast model, the ratio model cannot explain why different groupings of the faces arise in light of different preferences. Only the contrast model seems useful for explaining the context-sensitivity associated with PC.

4.6.2. Far-fetched

The main thread of previous criticisms on the feature-matching model centres on the fact that the sets of discrete features that describe the objects are given. Decock and Douven (2011) have argued that this makes Tversky's approach circular. On behalf of Goodman (1972), they argue that a good theory of similarity can be used to define properties because similarity is the most fundamental relation between objects. In their (and Goodman's) view, a feature is a property. In contrast

[7] Note that the contrast model can also make absolute distinctions. If people prefer smiling faces over frowning ones, round faces over sharp faces and curved eyebrows over straight eyebrows, then $S(c',d') = 1 < S(a',b') = 2 < S(a,b) = 4 < S(c,d) = 6$. Correspondingly, no grouping is plausible since all pairs are differently similar.

[8] The same result obtains when the weights are ignored such that $S(\cdot,\cdot) = 1/1+1+1 = 1/3$.

to this view, Tversky seems to think that features come prior to similarity. He presupposes the domain of objects, Δ, where each object is represented by a set of features (section 4.2. This presupposition is problematic because it makes the theory of feature-matching viciously circular. Similarity cannot be used to define properties if similarity itself is defined in terms of a set of properties associated with the objects under comparison.

Another previous source of criticism comes from Shanon (1988), who argues that Tversky's model is explanatory poor because it leaves the discrete sets of features unspecified. Shanon calls this "the problem of feature specification" (Shanon, 1988, p. 309). He illustrates the problem with an analogy to a communication task. He asks: which features should a person, a, choose to describe to another person, b, what face she is talking about? There is an innumerable amount of features with which a could describe the face to b. Which are those relevant to both a and b? A fixed objective standard for evaluating which features should be specified seems necessary for making the communication successful. In other words, the problem is that Tversky's theory offers no criteria according to which the relevant features should be selected prior to the matching process. More generally, the problem of feature specification seems to be important for a variety of similarity-judgement problems. For instance, which are the features that are relevant to identify a cup? Is it the shape? If so, is handleness more important than roundness? What about the colour? Is it irrelevant for identifying a cup? The roundness of a cup may be more important in a classical pottery course, while it may become less important in a modern design outlet for cups, where cups may have a futuristic shape.

Taken together, both Decock and Douven's and Shanon's criticisms centre on the problem that Tversky assumes either similarity or features as given. My own worries with the feature-matching model expand on these previous criticisms and focus on the explanation of directionality. These worries concern three assumptions: firstly, the implicit assumptions that the process of feature-matching seems to rely on symbolic thought (section 4.3.1), secondly, the assumption that directionality relies on an analysis of the semantic roles of the objects in a comparison (section 4.3.2), and thirdly, the explicit assumption that objects are decomposable into features (section 4.2.1).

With regards to the first and second assumptions, my worry is that the feature-matching model seems to be limited to an explanation of directionality in symbolic similarity judgements. Intuitively, these judgements seem to require two ingredients: symbolic thought and a prior semantic analysis of the objects' referents in a similarity statement. To give some background on this intuition, I have argued that it is implicit in Tversky's theory that subjects' internal comparisons take the form of beliefs or propositions (section 4.3.1). For instance, judgements of the similarity between a and b have the same systematic structure as statements of the form 'a is like b'. What is more, I have explained that a similarity judgement becomes directional when the relative salience of the distinct sets of features associated with a and b changes. In the explanation, the change in saliency depends on prior knowledge about the semantic roles of the

objects' referents in a similarity statement (section 4.3.2). For example, Tel Aviv is more similar to New York than vice versa because the distinct features of the more important object (i.e., New York) become more salient when New York is the object rather than the subject in a statement like 'Tel Aviv is similar to New York' (or vice versa). Following the contrast model, the increased salience of New York's distinct features makes the objects less similar. This explanation accommodates a wide range of empirical data (Tversky, 1977; Tversky & Gati, 1978) but its underlying assumptions are problematic.

The first assumption is quite controversial. One view, defended by Barsalou (2008), is that also amodal symbols (e.g., the signal '.—.') must be grounded in modal qualities perceived by the system (e.g., in auditory qualities). That is, the content of these symbols must come from perception. Tversky's assumption that features (e.g., the signal's temporal length) are discrete is at odds with this qualitative description because perception is inherently continuous (e.g., there seems to be no clear perceptual boundary between a crimson- and a scarlet-red colour shade). It is difficult to explain the perception of similarities with the feature-matching theory of PS because perception is at least sometimes continuous but in Tversky's theory, features are always discrete.

The second assumption is problematic when it comes to explanations of directionality in simple perceptual similarity-judgement tasks. In such tasks, the subject-object relationship between the objects under comparison might not be known to the subject. The finding of directionality in the Morse Code data seems to allow for this possibility; in Rothkopf's (1957) experiment, subjects were unfamiliar with the international Morse Code alphabet. It is unlikely that they could have identified the subject-object relationships between the signals. This raises further questions, such as: according to the feature-matching model, would children and animals who cannot identify subject-object relationships be capable of representing the similarity between pairs of objects? Tversky's explanation of PS as a function of feature-matching is difficult to generalise across different types of species, stimuli and modalities. It seems to be limited to comparisons of objects that are semantically analysable. This may motivate arguments in favour of Shepard's (1987) competing theory of PS. The assumption that PS is a symbolic process is generally difficult. One view, defended by Barsalou (2008), is that also amodal symbols (e.g., the signal '.—.') must be grounded in modal qualities perceived by the system (e.g., in auditory qualities). In other words, the content of these symbols must come from perception. Tversky's assumption that features (e.g., the signal's temporal length) are discrete is at odds with this qualitative description because perception is inherently continuous.

There are two problems with the third assumption (that objects can be decomposed into features). The first problem is that this assumption is difficult to combine with psychophysical models of how the perception of features works. The second problem is that this assumption seems to fall short on some evidence from developmental studies. I start with the first problem. Consider an analogy with the geometric model, in which dimensions shall play the role of features. In

psychophysical models of perceptual-object representations, it is common to distinguish between 'inseparable' and 'separable' dimensions (cf. Cheng & Pachella, 1984; Gärdenfors, 2000; Maddox, 1992; Melara, 1992; Shepard, 1994). Accordingly, a set of dimensions is inseparable if an object that is assigned a value on one of the dimensions must also be assigned a value along all the other dimensions that it is integrated with. For example, the dimensions hue, saturation and brightness in the colour space are integral because it is impossible to represent a colour shade along only one of them (e.g., only the hue dimension); representation of colour requires the simultaneous assignment of values along the hue, brightness and saturation dimensions. Conversely, a set of dimensions is separable if it is possible to represent an object's property by assignment of values on a single dimension in this set without also assigning it values on the other dimensions. For instance, the shape of a chair can be represented independently of the chair's colour. The analogy suggests that some psychological representations of perceptual objects, particularly psychological representations of properties such as colours, require a structure that is inseparable. If such representations of objects are comparable to dimensions, then some psychological representations of objects will not be decomposable. In Tversky's model, the object-base, Δ, is decomposable into discrete sets of features. The assumption that features correspond to psychological representations of properties of objects seems to be compatible with the assumption that psychological representations are separable but seems to clash with the assumption that some psychological representations of objects are inseparable. As argued above, the latter assumption is basic to at least some psychophysical studies of object perception. Because it is incompatible with this assumption, Tversky's model cannot take into account the results of experiments in such studies. Thus, Tversky's model seems to be disconnected from some of the evidence on the perception of similarities between objects and limited to an account of PS processes that involve only separable object representations.

I now turn to the second problem. Some findings in developmental studies suggest that psychological representations of objects in children are not (yet) decomposable. It is common to find that young children below the age of five typically confuse the height of a liquid in a container with the liquid's volume and just with time (typically after the age of five) do they learn to distinguish between height and volume. However, at this stage, children are able to identify what colour the liquid has.[9] According to this view, the decomposability of objects (e.g., liquids) into discrete features (e.g., height and volume) is not cognitively given and needs

[9]Initially, these studies were used to test Piaget's (1964, p. 177) theory of conservation, which hypothesises that young children fail to understand conservation (e.g., conservation of volume when pouring a liquid from a wider into a narrower container). One explanation of this learning effect has been proposed by Gärdenfors (2000, p. 28). He explains the shift in children's understanding how liquids should be represented by arguing that children *learn* to represent the qualities along distinct dimensions. Leach (1964, p. 34) puts this trajectory nicely, and as follows: "... the physical and social environment of a young child is perceived as a continuum. It does not contain any intrinsically separate 'things.' The child, in due course, is taught to impose upon this environment a kind of discriminating grid which serves to distinguish the world as being composed of a large number of separate things, each labelled with a name."

further explanation.

Taken together, this section has argued that the feature-matching theory of PS is context-sensitive (section 4.6.1). I have argued that this aspect of the theory is helpful to explain why PC might vary from context to context. However, this is only possible with the contrast model and not the ratio model. I have subsequently supported the claim that Tversky's feature-matching theory of PS is too far-fetched (section 4.6.2). I have reviewed Decock and Douven's (2011) and Shanon's (1988) earlier criticisms. I have then outlined four additional possible limitations of Tversky's model. The first limitation was that under the assumption that similarity judgements take the form of similarity statements, the processes of feature-matching in Tverky's model seem to correspond to symbolic-thought processes, which seem difficult to account for the perception of similarities. The second problem was that Tversky's explanation of directionality relies on the assumption that subjects can assess the subject-object relationships between objects in a similarity statements, but subjects often cannot do this although intuitively, they can perceive the similarities between the objects. The third problem was that the assumption that features are decomposable is difficult to combine with psychophysical studies on the perception of integral object representations and studies that suggest that the ability to decompose features is not developmentally given. Overall, I suggest on this basis that Tversky's theory can accommodate much empirical data, but its assumptions are in many respects too idealised or far-fetched.

4.7. Conclusion

In summary, this chapter has proposed Tversky's (1977) set-theoretic theory of PS as an alternative to Shepard's (1987) geometric theory of PS. The key assumption of Tversky's feature-matching model is that PS is a linear function of the common and distinct sets of discrete features. I have explained this assumption in detail (section 4.2). I have explained how two versions of the feature-matching model each accommodate the directionality in judgements of similarity (section 4.3). I have illustrated this with Rothkopf's (1957) Morse Code data (section 4.4), which was already subject to my illustration of the method of MDS in chapter 3. I have evaluated Tversky's set-theoretic theory of PS in section 4.6. I have argued that, on the one hand, the theory is very flexible; in conjunction with the diagnosticity assumption, Tversky's feature-matching model offers one possible explanation of why PC is inherently context-sensitive. On the other hand, the assumption of discrete features is possibly too far-fetched.

Before closing this chapter, let me add a critical remark to my evaluation of Tversky's theory of PS. My exposition has illustrated that the theory has obtained a wide variety of empirical support and can accommodate many aspects of PS-judgement behaviour that were previously difficult to explain with the geometric

conception. However, Tversky's theory does not explain why the behaviour associated with PS takes the form that it does. For example, why is the PS between a and b sometimes different from the PS between b and a? This lack of explanation is problematic from the perspective of a reverse-engineering approach to PC (chapter 3). Recapitulate that this approach starts with a computational level analysis of what the problem of PC is and why the problem is appropriate (Marr, 1982). The logic of the solution to the problem is then evaluated with respect to its optimality given certain environmental conditions (Anderson & Matessa, 1990). Models following this approach are typically appreciated for their predictive power (Zednik & Jäkel, 2016, p. 666). Neither of these aspects are present in Tversky's theory of PS. The feature-matching model is different from a computational level description. It describes what the function is that the system has to compute but not why this function is rational or why it should be computed (section 4.2). The theory is descriptive but not normative. From a perspective other than data-fitting, it is still unclear why PS should be regarded as a function of matching distinct sets of features. There is no reason to think that feature-matching is optimal with regards to the agent's environment. So why think that the feature-matching function is a plausible model of a subject's representation of similarity? Together, these points suggest that Tversky's model of feature-matching hardly fits into a reverse-inference explanation of PC behaviour.

Taken together, the feature-matching model targets explanations of different cases than the geometric model of PS. It is not immediately clear in how far these models compete with each other. The directionality associated with PS judgements like the case in which Tel Aviv is more similar to New York than the other way around and the context-sensitivity in categorisations of sets of countries like Jamaica and Cuba against Russia, on the one hand, and Cuba and Russa against Jamaica, on the other hand, are a case in point; these cases are not in the domain of explanatory targets of Shepard's model of ULG. The next chapter compared and contrasts these two theories of PS and argues that they compete with each other.

5. Interim conclusion: The Shepard-Tversky debate

Both Shepard's (1987) and Tversky's (1977) models offer possible approaches to the problem of modelling PC. Both propose that the ability to recognise or generalise and classify or distinguish objects depends somehow on the internal representation of how similar the objects are. Although many of Tversky's empirical examples go beyond purely perceptual cases (such as in the countries example from section 4.5), I have argued that some of these cases deserve to be interpreted as genuine examples of PC. A case in point are the faces (section 4.5) and Morse Code examples (section 4.4). The question that both theories seem to address is: How is it possible for an organism to generalise its behaviour from one (previous) experience to another (novel) experience, even if these experiences are unique? For instance, upon having eaten a (portobello) mushroom, x, and encountering a (fly agaric) mushroom, y, should you eat y as well?

The standard solution to this decision task is to say that the organism uses a measure of the similarity amongst the experiences of x and y, or a measure of their difference, to decide. Roughly, the rule is: if the two experiences are similar, then generalise. If the two experiences are different, then don't. The challenge for both Shepard and Tversky is to answer the question of what 'similar (or different)' means. In some sense, x and y are similar: they both have a body and a head. In some sense, they are different: x has a brown colour, y is red with white spots.

In chapters 3 and 4, I have illustrated that Shepard's and Tversky's models offer different approaches to a more precise notion of PS. This chapter compares these approaches with respect to their differences and commonalities and outlines the theoretical debate that has emerged between them.

5.1. Conflicting assumptions about similarity spaces

The first difference concerns assumptions about the PS spaces. Shepard (1962; 1987) assumes that there is a PS space and defines PS as the inverse of geometric distance in that space. The key points of this approach are (i) that PS must comply with the axioms of geometric distance and (ii) that this definition of PS accounts for the exponential gradient of generalisation, which is described by the ULG.

Regarding (i), particularly the symmetry axiom stands out when contrasting Shepard's and Tversky's models. Recall that symmetry says that $\delta(x, y) = \delta(y, x)$, where δ stands for 'dissimilarity' (section 3.4.1). The underlying assumption of the geometric model of PS is that dissimilarity corresponds to an internal representation of stimulus difference and is directly proportional to a metric measure of distance. Thus, the symmetry axiom indirectly says that the PS between x and y must be the same as the PS between y and x. Tversky's (1977) alternative definition of PS refrains from the metric axioms. The motivation for this is the empirical observation that judgements of similarity are sometimes directional. For example, people judge Tel Aviv to be more similar to New York than vice versa. I have explained that Tversky (1977); Tversky and Gati (1978) interpret these effects as violations of the metric axioms because they implicitly assume that similarity judgements take the form of similarity statements (section 4.3.1). Accordingly, while the symmetry axiom says that $S(a, b) = S(b, a)$, this finding illustrates that at least in some cases, $S(a, b) < S(b, a)$ (or vice versa), where S stands for 'similarity', 'a' stands for New York and 'b' stands for Tel Aviv. Tversky's model is not violated by observations of directionality in similarity statements, if these are understood as effects of asymmetries in internal similarity judgements.

(ii) is a key data point in favour of the geometric model but not the feature-matching model. The ULG says that generalisation maps monotonically onto PS; the higher the measured degree of PS between two objects, x and y, the *exponentially* more probable it is to observe an organism to generalise behaviour from one to the other. In contrast, the feature-matching model defines PS as a *linear* combination of the sets of shared and distinct features that the organism associates with x and y. Tversky's key assumptions are that sets of features are (1) discrete and (2) chosen from a database of objects that can be decomposed into features.

Taken together, these definitions of PS seem to rely on conflicting mathematical assumptions. The claim that PS is a function of geometric distance relies on the metric axioms and the respective function has an exponential shape. The claim that PS is a function of matching sets of discrete features relies on axioms of set-theory and the respective function is linear. The axiom of symmetry cannot be combined with the asymmetric relationship between some sets in some applications of the feature-matching model, and the exponential shape of generalisation cannot be combined with the linearity of the feature-matching function. With respect to these mathematical assumptions, the models seem to be in conflict.

5.2. Different explanatory targets

The second difference is that the geometric and the feature-matching model seem to target different empirical phenomena. When contrasting their analyses of the Morse code data, the set of directionality data that Tversky had analysed (section 4.4) is only a subset of the data that is subject to the MDS solution

proposed by Shepard (section 3.3.2). This may be a symptom of different goals: while Shepard seemed to have aimed at a visualisation and explanation of how subjects represent clusters of pairs of signals (section 3.5), Tversky seemed to have aimed at an explanation of directionality effects and a confirmation of his focusing hypothesis (section 4.3).

These goals seem to clash with respect to the empirical observation that similarity judgements are sometimes directional (see section above). On the one hand, the geometric model fails to explain such directionality effects because there is no parameter in Shepard's geometric model that could increase or decrease the relative distance of two points when these are represented in the same geometric space. What is more, if PS judgements have to rely on the symmetry axiom, then these judgements cannot be asymmetric. If asymmetries correspond to directionality effects, then if the symmetry axiom would accurately describe how people judge similarities, we should not expect people to display directionality in how they judge similarities.

On the other hand, I have illustrated that the contrast model can accommodate the effect of directionality in the example of Tel Aviv and New York (section 4.3.3). To recapitulate, the model assigns the set of features distinct to Tel Aviv relatively less weight than the features distinct to New York when New York is the object and Tel Aviv is the subject. Conversely, when Tel Aviv is the object and New York is the subject, the parameters in the model are adjusted in the opposite way. Therefore, a relatively smaller set of distinct features is subtracted when New York is the object than when Tel Aviv is the object in the comparison. Note that the feature-matching model can also accommodate non-directional results. Symmetry holds if and only if the weighted cardinalities of the sets of distinct features are equal, and when these sets subtract equal amounts from the set of shared features in the contrast model.

5.2.1. Existing geometric solutions to the problem of directionality

In this section, I review two existing generalisations of Shepard's (1962; 1987) geometric model of PS that can possibly accommodate directionality effects. On this basis, I suggest that directionality is less problematic for the geometric model than Tversky had originally thought.

The first generalisation of Shepard's geometric model is Nosofsky's (1986; 1991) approach to a biased-choice model. In the model, the dissimilarity between two objects is represented as a weighted geometric distance (e.g., City-Bock or Euclidean distance, chapter 3) between them[1]. A change in the distance between a

[1] A simplified version of how Nosofsky (1986) expresses this is the following function: $d_{ij} = \left[\sum w_k |x_{ik} - x_{jk}|^r\right]^{1/r}$ (Nosofsky, 1986, 41, eq. 6), where w_k is the weight, w, attached to a dimension, k, and $0 \leq w_k; \sum w_k = 1$. This function expresses the dissimilarity between two objects i and j as the sum of the weighted distances betwen i and j along each dimension.

pair of objects is introduced by adjusting the weights attached to the dimensions along which the objects are represented. In the model, a greater (or smaller) weight represents that a subject pays more (or less) attention to the respective dimension. This leads to a shrinking or expanding of the metric scale of distances (which is otherwise equal along different dimensions) and can modify the metric distance function. Increasing the weight stretches the dimension wile decreasing the weight shrinks it. This modification of a metric distance function can express the directionality of a task; an initial distance between two points can be lengthened or shortened if a change in the direction of the task increases or decreases the weights assigned to the dimensions of the relevant space.

Nosofsky's approach originates from Shepard's (1957) model for predicting identification/confusion probabilities in generalisation and similarity-judgement tasks. This model is now known as the 'similarity-choice model' (Luce, Bush, & Galanter, 1963). In the similarity-choice model, PS is interpreted as a function of the conditional probability of giving an identical response to a test stimulus j as was previously given to a training stimulus i. The conditional probability is defined as the ratio of the bias associated with j to the distance between i and j in a multi-dimensional space.[2]

In analysing the similarity-choice model, Nosofsky (1991) has argued that the observed effects of directionality in data-sets of confusion-probabilities (such as in Rothkopf's initial data matrix, figure 3.2, left) are effects of cognitive biases that influence the perception of individual objects. On Nosofsky's analysis, the key to using the similarity-choice model to explain observations of directionality effects is a separation between PS and bias in the model. Nosofsky distinguishes the two as follows. PS is a metric relation between two objects while bias is "a characteristic pertaining to an individual object" (ibid., 94). Building on the similarity-choice model, Nosofsky shows how stimulus bias can be added to a multi-dimensional scaling solution of identification/confusion-probability data to derive cases in which the spatial relation between objects in the solution can become asymmetric. On this account, it is the differences in the biases associated with whatever object plays the role of the test stimulus, j, that account for differences in the identification/confusion probabilities associated with two objects a and b, while otherwise, the distance between a and b would be constant. If the bias associated with $a = j$ in one direction has a larger bias than $b = j$ in the other direction, then it is more probable to identify/confuse a with b than to identify/confuse b with a. On this basis, a geometric model can be used to predict asymmetries in the identification/confusion data; if two objects are associated with different biases, then, depending on the relative order in which they are compared to each other, their PS is likely to be different.

[2]Formally, the function is this: $pr(R_j|S_i) = b_j s_{ij} / \sum_k b_k s_{ik}$, where $b_j, s_{ij} \geq 0$ and $s_{ij} = s_{ji}$. R_j stands for a response (e.g., a rating on a Likert-scale) attached to a test stimulus, j, and S_i stands for a training stimulus, i. b_j represents the bias attached to j and s_{ij} represents the similarity between i and j. The corresponding model says that the probability of a response to the test stimulus given a training stimulus is a function of stimuli-similarity as well as the bias attached to the test-stimulus.

A special case of the similarity-choice model is Krumhansl's (1978) distance-density model. In Krumhansl's model, PS is defined as the geometric distance between a pair of points in a multi-dimensional space, while density is defined as a function of the distance number of points surrounding a point. The model proposes a new interpretation of similarity-judgement data, which is in terms of a function of PS that represents also information about how well a subject can discriminate two points in relation to their respective environments. Correspondingly, the new distance-density measure of PS is a function of the geometric distance between the points, a and b, and the sum of their weighted densities. Formally: $s(a,b) = f[\bar{d} = d(a,b) + \alpha\delta(x) + \beta\delta(y)]$, where $\delta, \alpha, \beta \geq 0$. $\delta(a)$ represents the density associated with the surrounding of a and $\delta(b)$ represents the density associated with the surrounding of b, $\delta(b)$, while α, β represent weights on these densities. On the distance-density model, the PS between two objects is a linear function of the distance between the object representations in geometric space and the density of the environments associated with each of the objects. In contrast to geometric distance, density need not satisfy the metric axioms. PS need not be symmetric because $\bar{d}(a,b)$ can be different from $\bar{d}(b,a)$. In particular, if $\alpha \neq \beta$, then $\bar{d}(a,b) \neq \bar{d}(b,a)$ if and only if $\delta(a) \neq \delta(b)$ (Krumhansl, 1978, p. 447). That is, both the densities and associated weights must be unequal to derive directionality. However, the distance-density measure of PS becomes symmetric when $\alpha = \beta$ while, simultaneously, $\delta(a) = \delta(b)$. If either of these conditions is not met, then $\bar{d}$ will be asymmetric.

Taking stock, this section has outlined two existing generalisations of Shepard's geometric model of PS. These are Nosofsky's (1986; 1991) approach to the similarity-choice model and Krumhansl's (1978) distance-density model. In the next section, I compare and contrast these approaches to how the feature-matching model derives directionality effects. I argue that it is possible that the biases in Nosofsky's and Krumhansl's solutions are different from those in Tversky's solution.

5.2.2. Comparison to Tversky's solution

According to Tversky (1977, p. 333), directionality can be expected if the task is directional way and if the focusing hypothesis ($\alpha \neq \beta$) holds. Under these conditions, $S(a,b) \neq S(b,a)$ if and only if $f(B) \neq f(A)$ (i.e., the distinct sets of features of the objects are unequal in their relative salience). For example, if the task is to judge the similarity between Tel Aviv and New York, or vice versa, and if more attention is paid to the distinct features of the subject than the referent of the comparison (i.e., $\alpha > \beta$), then the PS between New York and Tel Aviv should be lower than the PS between Tel Aviv and New York (i.e., $S(a,b) > S(b,a)$), since New York is (intuitively) associated with more distinct features than Tel Aviv (i.e., $f(B) > f(A)$).

In comparison to Tversky's account, Krumhansl's (1978, p. 453) explanation of how the distance-density model derives the desired asymmetries builds on the

following argument. Firstly, Krumhansl takes on Tversky's focusing hypothesis, that when a similarity-judgement task is directional, then the subject will pay more attention to one object than to the other (i.e., $\alpha \neq \beta$). Secondly, Krumhansl adds to this hypothesis the assumption that a difference in focus produces the following effect: the density of the environment surrounding one object will affect their judged PS more than the density of the environment surrounding the other object (or vice versa). However, it must be well noted that this effect will only be produced if and only if these densities are different to begin with. Krumhansl originally applies this explanation to the case of directional comparisons between two more or less prominent objects. Roughly, Krumhansl assumes that the prominent object shares many features with other objects while the unknown object shares only a few features with other objects. On this basis, Krumhansl argues that the point representing the prominent object in the model falls in a dense region in PS space, while the point associated with the unknown object falls in an isolated region in PS space.

When applied to the above example of cities, Krumhansl's explanation of the observed directionality would be that the task is directional, so that $\alpha \neq \beta$, while, at the same time, New York and Tel Aviv share differently many features with other countries. In particular, to produce the desired effect that Tel Aviv is more similar to New York than vice versa, the following must hold. If $\alpha > \beta$, then New York has to share more features with other cities than Tel Aviv does, so that New York is in a relatively dense region in PS in contrast to Tel Aviv. A possible problem with Krumhansl's explanations is that it is not clear why New York should share more features with other cities than Tel Aviv (e.g., smaller cities have different features, but not necessarily less features). This problem does not apply to Tversky's explanation, which only requires that the distinct sets of features associated with Tel Aviv and New York are unequally salient (but not necessarily unequal in number).

Nevertheless, whereas in Tversky's model, it is unclear what determines the attentional focus ($\alpha \neq \beta$), Krumhansl's model offers a bit more precise suggestion for how the notion of cognitive bias in a geometric model could be understood. PS is biased by density; a difference in the densities associated with the environments of each point accounts for a difference in the PS between them, when the order of the comparison is swapped. PS can be different between a and b or vice versa if these objects are located in relatively more or less dense subregions of the space. If a has a relatively dense environment and b has a less dense environment then $S(a, b) > S(b, a)$ (or vice versa). Thus, a difference in the PS associated with two objects is a result of a difference in the spatial density of their environments. Without regards to differences in the density, the initial geometric distance between a and b would be exactly the same. Krumhansl's (1978, p. 457) motivation for incorporating a bias as density is that, intuitively, within relatively dense subregions of PS space, finer discriminations amongst object representations are possible than within less dense subregions. This idea adds plausibility to the original similarity-choice model that was inspired by Shepard (1957). It emphasises the fact that PS is a 3-place predicate; PS not only depends

on dissimilarities associated with two objects that share properties but also dissimilarities associated with other objects that share these properties (i.e., objects in the environment).

Taken together, Nosofsky's analysis and Krumhansl's specification of the similarity-choice model illustrate that the geometric model of PS can be extended to accommodate directionality effects by adding a bias to the metric-distance function. However, this does not show that directionality in the geometric model is the same as directionality in the feature-matching model. It is possible that these types of models (i.e., geometric versus feature-matching) provide different solutions to the problem of directionality. The early solution to the problem of directionality that is offered by the similarity-choice model is compatible with a variety of different algorithms that could be used to compute the function of the conditional probability of giving an identical response to a test stimulus j as was previously given to a training stimulus i. In this respect, the abstract nature of the similarity-choice model seems to reflect a computational level approach to PS. The function does not specify what the cognitive bias is at the level of psychological process. It is possible that the PS-judgement process in a specific case of the similarity-choice model is different from the one in Tversky's (1977) contrast model (equation 4.2), even if both incorporate the abstract notion of cognitive bias. In the next section, I support this proposal by contrasting differences in the models' implicit assumptions about the structure of mental representations. Table 5.3 makes these differences explicit.

5.3. Different structures of mental representations

To interpret the differences in table 5.3 in a philosophical context, I position the geometric and feature-matching models alongside two camps in the literature on mental representations of objects and categories. Roughly, following the first camp, some mental representations, particularly those used in perception, have a continuous structure (Beck, 2019; Goodman, 1968; Haugeland, 1981). According to this camp, such representations have contents that often cannot be clearly differentiated from one another. For example, the perceptual experience associated with two blue colour shades allows for more fine-grained or richer distinctions than a distinction that is based on the colour concepts turquoise and aquamarine. Following the second camp, all mental representations, or at least those that can carry semantic content, take the form of discrete symbols (Fodor, 1975; Fodor & Pylyshyn, 1988). For example, the mental representation of a cat takes the form of a symbolic structure, CAT that can figure in the belief 'the cat is on the mat.' Correspondingly, the propositional content of the belief is a function of the content of the individual symbols, e.g., CAT and MAT, and how they are composed to form this proposition.

Shepard's geometric model seems to follow the first camp in assuming that at least some object representations derive from the phenomenal experiences of per-

Geometric model	**Feature-matching model**
Representations of objects are vectors in geometric space.	Representations of sets of features are discrete and objects are decomposable.
The space consists of continuous dimensions. Partitions can be performed on them.	The basic elements are whole objects in the data base, Δ. Objects in Δ generate discrete sets of features.
PS is measured by geometric distance.	PS is measured by set-theoretic overlap.
The model partitions the continuous dimensions into sets of regions in geometric space.	The model divides the database with all objects into subsets with objects that increase similarity with respect to a diagnostic feature (cf. section 4.5).
A category's size is the area occupied by its region in PS space.	The size of a category is the cardinality of the set of overlapping features of objects in this category.
PS determines categorisation; a consequential region represents the average similarity of any object in the area covered by the region in geometric space.	The categorisation of objects according to their associated sets of common and distinct features happens prior to similarity processes.

Table 5.1.: A contrast between the implicit assumptions about the structure of mental representations in Shepard's (1987) geometric model (left column) and Tversky's (1977) feature-matching model (right column).

ceptual objects such as, for example, colours or tones. As can be seen in earlier work (Shepard, 1957, 1962, 1981/2017), Shepard's motivation for modelling PS as geometric distance originates from the idea that relations between mental representations must map onto relations between objects in the world. In this work, Shepard claims that mental states represent perceptual objects because they preserve structure, not because they resemble those objects like pictures. For him, the extensions of concepts or consequential regions (which, for him, are psychological representations of the kinds that these objects belong to) are the result of positive correlations on an evolutionary scale along the dimensions of the assumed psychological space (Shepard, 2001, 600).

On Shepard's account, structure-preservation is relevant for generalisation. A structure-preserving mapping allows the subject to represent invariances in the experimental stimuli (e.g., colour constancy in perception or the perception of the successive presentation of an object in two positions as it moving along a path that connects these positions), enabling the agent to treat stimuli as the

same or different. This coincides for Shepard with a treatment of the stimuli in a survival-conducive way. For example, mental representations of poisonous and edible mushrooms should be structured in such a way that they allow the agent to successfully distinguish poisonous mushrooms from edible ones because only in that way will the agent maximise her chances to eat the right kind of mushroom (e.g., the one that is good to eat).

On this basis, Shepard justifies the definition of PS as a geometric distance by the assumption that there exists a second-order isomorphism—a functional relationship between the invariances between objects in the world and the invariances between mental representations in the mind ('second-order', thus, because it describes a relationship between two relationships, (Shepard, 1981/2017, pp. 290-292)).

Empirical support for the assumption that the geometric-distance definition of PS is fruitful comes from an experimental study on mental rotation (Shepard & Metzler, 1971) and the perception of the shapes of states (Shepard & Chipman, 1970). In the former, Shepard and Metzler find evidence that the time it takes to internally rotate a representation correlates with the time it takes to rotate its referent stimulus, confirming the hypothesis that each step in the internal transformation of the representation corresponds to a step in the external transformation of the object. In the latter, Shepard and Chipman find that the cartographic features that relate pairs of shapes of U.S. states to a common category (e.g., features such as 'smallish & wiggly', 'vertical & irregular' etc.) are mirrored in the relations between a subject's perceptual representations of those shapes. The results of Shepard and Chipman's experiments show that pairs of the states' shapes that could be identified with properties such as 'smallish & wiggly' were on average judged across subjects to be more similar than, for instance, a shape that is wiggly and a shape that is irregular (Shepard & Chipman, 1970, p. 10). Shepard and Chipman take this to be evidence for the geometric model because, on the assumption that distances between objects are constrained by the metric axioms, it was possible to reconstruct patterns that had emerged from subjects' averaged data. In particular, it became possible to reveal that judgements indicating that some pairs were consistently judged to be more similar than other pairs corresponded to the classifications of the states' shapes according to the external relationships between their cartographic properties.

In contrast to Shepard's geometric model, Tversky's feature-matching model seems to follow the second camp. I have argued in section 4.6.2 that Tversky implicitly assumes that PS correspond to processes of symbolic thinking in so far as features in the model play the roles of atomic and amodal structures. Following the feature-matching model, one way in which PS can be understood is as a process of classification of objects into those that have a target feature and those that do not have the feature. These classifications represent disjoint classes of objects. An example is given by Sattath and Tversky (1987), who use the feature-matching model to study the relationship between object classification and object similarity. They write:

> the predicate 'two legged' can be viewed as a feature that describes some animals; it can also be seen as a class consisting of all animals that have two legs. The relation between a feature and the corresponding cluster is essentially that between the *intension* (i.e., the meaning) of a concept [e.g., TWO LEGGED] and its *extension* (i.e., the set of objects to which it applies [e.g., the set of animals that are two legged]) (Sattath & Tversky, 1987, p. 16, original emphasis)

However, the feature 'two legged', if discrete, cannot overlap with other features (e.g., 'being tall' and 'being alive'). Thus, on this account, if a feature such as 'two legged' represents the meaning of a concept, then the account does have a difficulty to specify how the contents of concepts can partly overlap. In a space of discrete sets, it is not clear how the model represents that only a subset of the set of objects that are two legged is simultaneously alive. On this account, features are themselves like symbols and do not naturally overlap in a meaningful way. To make the symbols overlap meaningfully, an additional function to interpret their relations is required.

5.4. Different interpretations of the data

How do these assumptions fit together into an overall picture of the study of PS? One idea, mentioned in Shepard and Arabie (1979, p. 89) and Shepard (1981/2017, pp. 395-396) is that different sets of representational assumptions fit better for modelling different types of psychological data. For example, when modelling relations among stimulus data such as individual words, discrete representations of features seem better than a low-dimensional continuous space. In contrast, a continuous-spaces solution seems to provide the most natural representation of perceptual stimuli such as colour shades. Shepard and Arabie (1979) argue that in any of these cases, the different formats of representation make it relatively easy for the modeller to interpret relations between the data. Shepard (1980) and Shepard and Arabie (1979) make this more explicit in their comparisons of spatial and hierarchical methods[3] of representing the same similarity-judgement data (e.g., as collected from the same data matrices).

They focus on data representing similarity judgements of pairs of vowel phonemes. Shepard (1980, p. 397) argues that the continuous, spatial representation preserves the parallel orderings of the voiceless and the voiced fricatives [...] with respect to place of articulation and, hence, represents such facts as that the middle fricatives are more often confused than the extreme fricatives." He contrasts this with a discrete-clustering representation of the same data sets. Accordingly, only the discrete representation "reveals that [...] place of articulation [is] more salient than presence or absence of affrication for voiced consonants, while for the voiceless consonants absence of affrication [becomes] more salient owing to

[3]They particularly focus on the different methods of multi-dimensional scaling, tree-fitting and hierarchical clustering.

the correlated presence of an initial burst [...]" (Shepard, 1980, p. 397). In contrast, it seems that the spatial representation of the vowel-phoneme data often maintains information that gets lost in a hierarchical model that uses discrete representations of the data. In the case of vowel phonemes, the lost information is the possible overlap of groups of phonemes. Shepard and Arabie (1979, p. 91) argue that "once all the voiced consonants were grouped into one cluster (disjoint from the cluster of voiceless consonants), it would no longer be possible to group either all the stops or all the fricatives into one cluster." Taken together, Shepard's (1980) and Shepard and Arabie's (1979) analyses of the vowel-phoneme data suggests that those aspects of the data that contain possible overlaps in properties of vowel phonemes are easier to model with a spatial solution while aspects of distinct classifications of the vowel phonemes are easier to model by the discrete-hierarchical solution.

This suggests a constructive way of thinking about the Shepard-Tversky debate: the preferred definition of PS may depend on the modeller's intentions, that is, the intentions of Shepard's and Tversky's intentions. One way in which such intentions can be seen is in terms of Weisberg's (2012) idea that the interpretation of a model depends in part on the modeller's intended scope. According to Weisberg, the intended scope "specifies which aspects of potential target phenomena are intended to be represented by the model [...]" (Weisberg, 2012, pp.39–41). When transferring this characterisation to interpretations of the geometric and the feature-matching models, one option seems to be that Shepard and Tversky just have *different* intended scopes when they study PS in the context of generalisation and explicit judgement. For example, in the vowel-phoneme case, the geometric model and the feature-matching model can bring out different aspects of the same data by virtue of different representational methods and algorithms such as clustering or MDS algorithms. These different ways of representing the data may be more or less appropriate depending on the particular explanatory target (e.g., to group versus to distinguish objects).

Another idea is to view the geometric-spaces model of mental representation as a third way *alongside* symbolic and associationist approaches to modelling mental processes. This picture has been proposed by Gärdenfors (2000). Accordingly, the geometric approach complements these previous approaches, while providing a basis for connecting them. According to Gärdenfors, one reason for why the geometric approach is especially suited for this role is that it is flexible; the geometric model can accommodate both continuous and discrete representations. Gärdenfors (2000, p. 7) illustrates this with the example of a phylogenetic tree, in which each node represents a different species and the hierarchical arrangement of paths connecting them represents biological kinship relations. Older species are higher up in the tree and younger ones are lower down, so that the vertical axis implicitly represents the (continuous) dimension of time. It is possible to make gradual distinctions between species because the length of the path between any two nodes represents the phylogenetic difference between the species that they represent, relative to the lengths of other paths. Gärdenfors example for such relations is that "birds and reptiles are more closely related than rep-

tiles and crocodiles" (Gärdenfors, 2000, p. 8). At the same time, the tree has a discrete structure; the hierarchical arrangement of the nodes allows one to infer distinct kinship categories. Thus, the geometric model can encompass aspects of continuous as well as discrete representations and, although the assumptions about continuous and discrete representations are different, they can be related intuitively.

In sum, there is an obvious contrast in the explanatory targets and implicit assumptions about the structure of mental representations in the geometric model and in the feature-matching model. I have explained two possible interpretations of this contrast. The first interpretation was inspired by Shepard and Arabie (1979) and is that the decision about which model is appropriate to model different aspects of the data depends on which aspects of the data the modeller is interested in (i.e., the intended scope). Shepard and Arabie's (1979) analysis of solutions of the vowel phoneme data is an example. Here, if the modeller is interested in aspects of generalisation and recognition, she should use the geometric model. In contrast, if she is interested in aspects of classification and distinction, she should use the feature-matching model. The second interpretation was inspired by Gärdenfors (2000) and is that representations of different aspects of the data can be combined in the geometric model. Gärdenfors' example of a kinship tree illustrates that the assumptions of the geometric and feature-matching models do not entirely conflict with regards to the structure of their representations. But at the same time, both approaches to PS compete with respect to the explanation of how people may possibly judge similarities (section 5.2).

On the whole, this conglomerate picture calls for better organisation. Intuitively, it seems reasonable to connect aspects of the different representational structures because minds are presumably capable of using all of these structures when representing similarities, i.e., none should be discarded a priori. But it also seems reasonable to keep aspects of these models apart that concern different possible mechanisms. Is there a single approach that can connect aspects of these models of PS, while keeping aspects of them apart? The next section reformulates this question as a problem of unification.

5.5. A problem of unification

Modellers need a guiding framework that can relate these models to each other while keeping an account of the empirical data. Given the empirical support that both models have received, neither of them can be discarded. However, given their theoretical differences in the underlying axioms, their different explanatory targets and different assumptions about the structure of mental representations, it is a challenge to integrate them to a single coherent theory of PS that can possibly explain aspects of possible mechanisms underlying PC. Nevertheless, the foregoing analysis suggests that these distinct approaches to PS may possibly overlap in various ways. For instance, Tversky's theory can accommodate non-directional

(symmetric) similarity-judgement behaviour as well, so long as the focus on each object in the comparison is balanced (i.e., if $\alpha = \beta$). Likewise Shepard's model is capable of dealing with stimuli that can be analysed in terms of discrete objects, for instance, if the dimensions are discrete and the geometric space is carved up into sets of points, and can possibly be extended to accommodate directionality as well.

(How) can the divide between these approaches be bridged and the confusing picture be systematically organised? A sketch of this problem is illustrated in figure 5.1. The idea is that Shepard's and Tversky's models suggest two distinct explanans, 'similarity A' and 'similarity B', that are associated with different aspects of PC behaviour. The challenge is to unify the observations of the exponential gradient and the effect of directionality, despite the theoretical conflict between Shepard's and Tversky's approaches.

In Part II of this thesis, I approach this problem from a Bayesian perspective on concept learning and propose a unifying approach to PC. The motivation of this approach is to better comprehend the diversity of empirical phenomena associated with PS. My approach builds on Tenenbaum & Griffiths' (2001) Bayesian model of concept learning and focuses on combining the predictive powers of Shepard's and Tversky's models (indicated by the intersection in figure 5.1) while allowing them to compete with each other (indicated by the disjunction in figure 5.1). I will discuss Tenenbaum & Griffiths' model in the next two chapters.

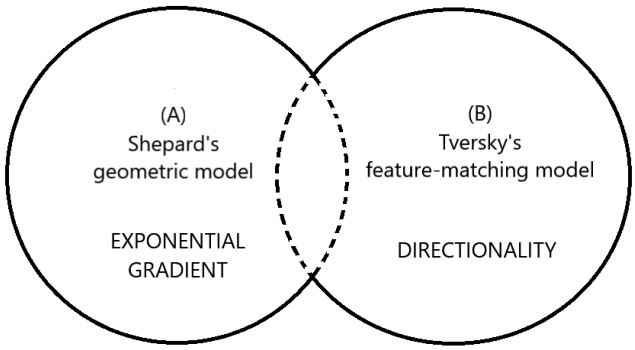


Figure 5.1.: Depiction of the relationship between Shepard's (1962, 1987) geometric model of similarity and Tversky's (1977) feature-matching model of similarity. The two accounts are presented as two partly overlapping sets of statements that describe two distinct types of psychological similarities on the basis of their empirical predictions. A represents the geometric model and predicts effects of exponential generalisation. B stands for the feature-matching model and predicts directionality effects. The area enclosed by the dashed lines describes an area of a possible overlap of the predictions of these models.

Part II.

Bayesian inference

"The best-developed mathematical tool for dealing with inference under uncertainty is probability theory." (Zednik & Jäkel, 2016, p. 3955)

6. A Bayesian approach to perceptual categorisation

6.1. Introduction

This chapter argues that PC can be analysed as a Bayesian-inference problem of generalisation at the computational level of explanation. This argument is inspired by Tenenbaum and Griffiths' (2001, henceforth T&G) Bayesian inference model of similarity and generalisation, an explication of which is at the heart of this chapter. I explicate T&G's claim that Bayesian inference explains psychological similarity (PS) and generalisation and argue that T&G's Bayesian model provides a possible alternative to similarity-based explanations of perceptual categorisation (PC), which have multiple facets that do not fit together well.

Section 6.2 presents 2 examples of a Bayesian analysis of PC, which are inspired by T&G's Bayesian model of concept learning. Section 6.2.1 presents a Bayesian analysis of PC, which is inspired by T&G's revision of Shepard's (1987) ideal generalisation problem. Perceptual categorisation, on this view, is the same as the problem of judging similarities or generalising from one object to another. Section 6.3 systematically investigates the key ingredients of a Bayesian analysis of PC at the computational level of explanation. Section 6.4 focuses on the ingredients for a generic solution to the problem of modelling PC. The solution is T&G's *size principle* of preferring smaller categories in the inference problem. Section 6.5 draws a strong connection between T&G's size principle and Shepard's geometric model, showing that the size principle implicitly relies on a representational structure of the relevant concepts (e.g., FLY AGRARIC MUSHROOM). Section 6.6 argues that the Bayesian approach to PC employs a reverse-engineering strategy to discovering the mechanisms underlying PC behaviour. I conclude that the Bayesian model provides an elegant perspective on PC that combines the approaches to PS and generalisation.

6.2. Bayesian-style thinking about categorisation

What does it mean to think of PC as a Bayesian inference? In their paper, T&G (2001, p. 629) mention several examples for interpreting generalisation as a problem of Bayesian inference. I will focus on 2 of these to make the conception of PC as a Bayesian inference intuitively accessible.

BABY ROBINS

A baby robin needs to learn to distinguish edible from inedible worms on the basis of their different levels of skin pigmentation. Assume that the first worm given to the baby robin has a pigmentation level of 60. Given that the worm with pigmentation level 60 is edible, what other worms are good to eat? Is a worm with a pigmentation level of 47 edible? What about a worm with a pigmentation level of 80?

HORMONE LEVELS

A doctor has to determine whether a hormone, naturally produced by the body, affects a patient's health. A first blood analysis reveals that a patient, x, has a level of 60 of this hormone. x has been in the hospital for a long time and has recovered from their illness, so the doctor knows that x is healthy. They have to say for any new patient, y, whether y is healthy or unhealthy on the basis of blood tests that indicate y's hormone levels. Given the knowledge that 60 is healthy, is a patient with a level of 47 healthy? What about a patient with a level of 80?

In each of these Bayesian analyses of a categorisation task, the agent (e.g. doctor or baby robin) has to infer, on the basis of a set of available concepts (HEALTHY/UNHEALTHY or EDIBLE/POISONOUS), what the correct concept is in light of the observations, x and y (hormone levels or worms with different pigmentation levels) and the knowledge that x is an instance of one of the candidate concepts. This offers a preliminary definition. For any object x, which is known to be in a category, C, it has to be inferred how plausible it is that a new object, y, is also in C (or, alternatively, is not in C). The task of the corresponding inference is to identify the relevant category, C.

Let me add two points of clarification on these examples. Firstly, the approach assumes that in all examples, a candidate category or concept that represents the to-be-inferred properties is already available. When inferring whether the concept HEALTHY correctly represents a hormone level of 80, some representation of what it means to be healthy is at least implicitly needed to carry out the categorisation task (cf. Fodor, 1975). In other words, candidate categories need not be inferred from scratch.

Secondly, on the Bayesian approach, both the assignment of a category and the perception of the property that is represented by the category are distinct capabilities. Not every case of perception involves a category assignment and not every case of category assignment involves perception. The difference is that perception can be non-inferential, but category assignments are always inferential. For example, perceiving a worm with pigmentation level 80 does not require one to infer that this observation is an instance of the category 'good-to-eat' worms.

An extra inferential step would be needed to assign the instance to a category.[1] In what follows, I explicate my claim that PC can be seen as a problem of Bayesian inference on the basis of T&G's Bayesian model of concept learning.

6.2.1. Generalisation as a problem of Bayesian inference

Following Marr (1982, p. 22), the computational level of explanation of how an information-processing system solves a given task consists of two questions. One question is about *what* is being computed. The other is about *why* a task should be solved with a function of some type x and not one of some type y. The point of the analysis of the information-processing task at the computational level is to identify the rationale underlying the problem that the information-processing system is presented with.

When PC is regarded as an information-processing problem, the analysis focuses on two aspects. On the one hand, this method analyses the task of PC as a Bayesian inference problem and a generic function that solves the problem. This explains what is being computed. On the other hand, the method analyses the agent's environmental conditions with a normative standard of how they have to be met. This explains why the generic solution takes the form that it does. From the perspective of a Bayesian analysis of the task of PC, an agent is rational if and only if she solves the inference problem in the hormone levels and baby robin examples (to infer the relevant category from the perceptual observations) by computing the relative probabilities for any candidate category in light of the available evidence by following Bayes' Theorem. The 'correct' category may be either a single category that obtains the highest probability value or the average of a set of candidate categories with the highest probabilities. I will now explain this based on T&G's (2001) Bayesian model of concept learning.

In their model, T&G assume a Bayesian task analysis of generalisation. They assume that generalisation is a problem of Bayesian inference. According to T&G, the problem is this:

> [G]iven an encounter with a single stimulus (a patient, a worm) that can be represented as a point in some psychological space (a hormone level or pigmentation level of 60), and that has been found to have some particular consequence (healthy, good to eat), what other stimuli in that space should be expected to have the same consequence? (Tenenbaum & Griffiths, 2001, p. 629)

T&G's Bayesian view on the problem is basically a more abstract version of Shepard's ideal generalisation problem. To see this, compare T&G's framing with Shepard's original formulation of the problem (Shepard, 1987, 1994):

[1] In analogy, when explaining how someone infers from the meaning of the concept DEMOCRACY the meaning of a concept like EQUALITY, no reference to perceptual capabilities is necessary, as the inference seems to involve only highly abstract concepts.

> The problem that a positive or negative encounter with an unfamiliar object poses for an individual is just the problem of inferring the consequential region to which that object belongs. A bird that ingested a caterpillar bearing particular coloration and markings and found it detectable or sickening, must decide whether another object of more or less similar visual appearance is of the same natural kind and should therefore be seized or avoided, respectively. Generalization is thus a cognitive act, not merely a failure of sensory discrimination. Indeed, an animal would be ill served by the assumption that just because it can detect a difference between the present and a previous situation, what it learned about that previous situation has no bearing on the present one. (Shepard, 1987, p. 1319)

By a 'cognitive act', Shepard means the mental process of testing the hypothesis that the second object, of "similar visual appearance" is of the same natural kind as the first object (the caterpillar).

On both accounts, generalisation implies an assignment of each object to a concept (on T&G's approach) or a consequential region (on Shepard's approach) in an agent's psychological space. What is different is that it is implicit in the second quote that the structure of this space relies on evolutionary norms. Shepard's position is that generalisation is an inference process that is optimised considering the agent's environment. Under this position, the agent should assume that there could be an underlying connection between the first and the second object despite their apparent differences. The agent should assume that there exist natural kinds that the objects could possibly belong to. Why exactly this assumption is plausible is not clear from this quote, but in other work, Shepard has argued that organisms have adapted their internal psychological mechanisms that govern generalisation behaviour to the structure of natural kinds in the world (Shepard, 1981/2017)[2]. Shepard's framing of the problem as an adaptively optimised inference commits his analysis to answers about the representational structure of the agent's psychological space. This is because the agent has to somehow represent the assumed natural kinds. In contrast, T&G's version disregards such assumptions about evolutionary norms and does not explicitly commit to the assumption that the inference must be optimised given the agents' natural environment. This is one reason for why T&G can express some agnosticism in the first quote about how consequential regions shall be individuated in psychological space.

Is it useful to stay neutral on the structure of the psychological space? One reason for why the neutral attitude could be useful is because it allows us to

[2]The key aspect of this work is Shepard's assumption that the structure of the agent's internal representations reflects the structure of natural kinds. The support for this assumption is that the organisational structure of the agent's mental representations must be such that it serves the agent's adaptive success. Shepard also assumes that the (physical) structure of natural kinds is best modelled in a geometric space. His assumption that there exists a structural mapping between the agent's internal representations and the structure of natural kinds supports his assumption that the structure of the agent's internal representations should be modelled in a psychological space with geometric dimensions.

not buy into the Shepard-Tversky debate. Consequential-regions talk is plausible only if one can assume that representations in the inference involve geometric structure. It is not trivial that the psychological space has a geometric structure. Tversky (1977) has illustrated this with evidence against the symmetry axiom (chapter 4). Getting rid of the assumption of geometric structure allows one to express the problem of generalisation on the basis of concepts with an unspecified representational structure.

Taken together the generalisation problem can be reconstructed considering the reverse-inference/engineering approach outlined in chapter 1. Correspondingly, Tenenbaum and Griffiths' (2001) Bayesian approach to concept learning has the following argument structure.

Argument structure of T&G's approach:

A. When psychological process, CL, is recruited by a Bayesian-inference task, some pattern of generalisation behaviour, E, is likely to be found.

B. In Bayesian-inference task T, E was found.

C. Hence, psychological process, CL, was recruited by Bayesian-inference task T.

Where 'CL' refers to a process of concept learning (e.g., the process of inferring a concept such as EDIBLE from examples x and y). T&G's theory explains concept learning on the basis of their Bayesian model of generalisation. The next section shows how T&G analyse generalisation as an inference problem at the computational level. One of the main contributions of this analysis is that this abstract description is compatible with multiple interpretations of the structure of psychological space, particularly with the idea of overlapping concepts. The virtue of this approach is that it can unify PC (as will be argued in chapter 9).

6.3. Key ingredients of the Bayesian model

A HYPOTHESIS SPACE

The first ingredient of T&G's model is an abstract hypothesis space, $\mathcal{H}$. $\mathcal{H}$ represents the agent's innate background knowledge about which possible concepts could be assigned to observable objects. T&G assume about $\mathcal{H}$ that "[...] one and only one element of $\mathcal{H}$ is assumed to be the true [concept corresponding to some unknown consequence,] C (although the different candidate regions represented in H may overlap arbitrarily in the stimuli that they include)" (Tenenbaum & Griffiths, 2001, pp. 630-631). Let us unpack this assumption.

In T&G's model, $\mathcal{H}$ is a probability distribution over a set of mutually exclusive hypotheses. Mutual exclusivity means that if there are two hypotheses, h and h^*, then both cannot be absolutely true at the same time, so it cannot be that

simultaneously, $pr(h) = 1$ and $pr(h^*) = 1$. Either both are absolutely false or one of them is somewhat true and the other is somewhat false. Hypotheses have the form $h : x \in C \wedge x \in C^*$ or $h^* : x \notin C \wedge x \in C^*$, where C is a concept and C^* is a concept whose members include all members of C and some other members as well. For example, if C represents the concept EDIBLE MUSHROOM and C^* represents the concept MUSHROOM, then h says that x is an instance of both concepts while h^* says that x is an instance of the concept MUSHROOM but not of the concept EDIBLE MUSHROOM. Note that the concepts can overlap, despite the fact that hypotheses are mutually exclusive. In the example, the concept MUSHROOM includes the concept EDIBLE MUSHROOM. Roughly, in terms of their extension, the set of edible mushrooms is a subset of the set of mushrooms. However, the hypotheses still have different truth-conditions; if it is true that x is an instance of the concept MUSHROOM but not of the concept EDIBLE MUSHROOM, then it must be false that x is both an instance of the concept EDIBLE MUSHROOM and of the concept MUSHROOM.

Following T&G's assumption, $\mathcal{H}$ can be specified more explicitly on the basis of three abstract components. (1) A set of countable, possibly infinite, and mutually exclusive hypotheses, $h_i, ..., h_n, ...$, (2) a set of object representations (stimuli) that belong to sets of concepts (each concept corresponds to a consequence), and (3) a probability distribution. I briefly comment on each component in turn.

Hypotheses are statements about possibilities.[3] In T&G's model, a hypothesis represents the possibility that a stimulus is an instance of a candidate concept. For example, a hypothesis might represent the possibility that the object x, a fly agraric mushroom, is an instance of the concept FLY AGRARIC MUSHROOM. At the abstract level relevant for T&G's approach, hypotheses are simply functions that assign objects from the set of pieces of available evidence, $e : x, y, ...,$ to a candidate concept. In T&G's model, $\mathcal{H}$ is a partition of sets of hypotheses, $h_i, ..., h_n, ...$ A plausible assumption of their model seems to be that $\mathcal{H}$ only contains hypotheses that are consistent with e. The intuition is that if it is not a logical possibility that x is an example of the concept FLY AGRARIC MUSHROOM then the occurrence of x cannot inform the hypothesis that x is an instance of the concept FLY AGRARIC MUSHROOM.

In T&G's model, concepts are not individuated. Concepts are abstract sets of possible stimuli that could have the same consequence. T&G's abstract description of the hypothesis space contrasts with Shepard's model and framing of the generalisation problem in the previous section. On Shepard's account, the consequential region is 'located' in the agent's PS space Shepard (1987, p. 1319). We know from chapter 3 that Shepard proposes a geometric structure for this space. The dimensions (e.g., hue, saturation or brightness) of this space represent aspects of perceptual experience (e.g., how bright something looks). A consequential region is a mental representation in the agent's PS space. For Shepard,

[3] This is a typical understanding in Bayesian epistemology and philosophy of science (Hartmann & Sprenger, 2010; Sprenger & Hartmann, 2019).

this region corresponds to "a particular class—what philosophers term a 'natural kind"' (Shepard, 1987, p. 1319) (although he does not specify what natural kinds are supposed to be). In Shepard's jargon of behavioural consequences, a concept captures all possible objects that, upon observation, would obtain the same behavioural consequence. With regards to the geometric interpretation of PS space, a concept covers the area that represents the average distances between any points that could fall in that area, while each point represents a way in which the concept could be instantiated. I interpret the consequential region in Shepard's model to represent the intension of a concept. That is, the consequential region is the agent's representation of what things could possibly be such that they could be instances of the concept. For example, the intension of the concept RED represents all the possible ways in which the agent could perceive a red-shade.

The equation of consequential region and concept is plausible from an explanatory perspective. As explanatory entities in psychology, concepts are kinds of mental representations that explain behaviour. In Shepard's case, the consequential region fulfils the same explanatory job, it is part of the explanation of why generalisation has the shape that it does (see also chapter 3). I show in section 6.5 that, on the basis of this contrast, T&G's explanation of generalisation abstracts away from such details in Shepard's explanation.

Sets of stimuli with the same consequences (concepts) have two characteristics. Firstly, they represent the extension of concepts that figure into mutually-exclusive hypotheses. Secondly, while hypotheses are mutually exclusive, sets of stimuli can overlap with respect to a concept. The first characteristic contrasts sets of stimuli with concepts (consequential regions) in Shepard's PS space. Sets of stimuli are discrete, while consequential regions can be continuous (like the dimensions of PS space).

As an example of the second characteristic, consider the set of tigers and the set of pumas. On the one hand, sets of stimuli can overlap with respect to a concept. Both sets are in the extension of the same class: WILD CAT. In other words, assignment of the experiences of tigers or pumas to the class WILD CAT need not be mutually exclusive because membership of the class applies to both tigers and pumas. On the other hand, each set can be associated with different and differently plausible possibilities. For example, the hypothesis that the label 'wild cat' refers to the set of tigers but not to the set of pumas and the hypothesis that the label 'wild cat' refers to the set of tigers and pumas must be mutually exclusive; these hypotheses cannot be true simultaneously. The relation between the sets of stimuli and hypotheses in $\mathcal{H}$ is the following. At the level of stimulus representation, $\mathcal{H}$ partitions sets of stimuli into mutually exclusive alternatives. A hypothesis is a membership-statement of the form 'x is in class C', where x could be a tiger and C could be the concept WILD CAT. By definition of a partition, the truth-values of these hypotheses cannot overlap because the assignment of truth-values to hypotheses in $\mathcal{H}$ must be mutually exclusive. Thus, also on a set-

theoretic interpretation, the simultaneous existence of overlapping sets of stimuli and mutually exclusive hypotheses in T&G's model is not inconsistent.

The probability distribution represents the agent's knowledge as a conditional probability function, $pr(h|e)$. The function, $pr(h|e)$, assigns each hypothesis, $h \in \mathcal{H}$, a probability value. This assignment is conditional on the observation of an example, e, which is the evidence for h. In T&G's model, '$pr(h|e)$' represents the learner's degree of belief in h after observing e, that is, how strongly the learner believes that h is true, compared to all other hypotheses in $\mathcal{H}$, in light of e. The strength of this belief is represented by a probability value, a value in the interval $[0, 1]$.

A GENERALISATION FUNCTION

The second ingredient of T&G's model is a generalisation function, which formalises the Bayesian inference task.

The task is that the agent has to infer from two observations, x and y, whether these belong to the same concept (or whether they do not). The generalisation function recovers this task formally. The following is a modification of T&G's formulation to emphasize that the hypothesis is a statement about the concept. For their original formulation, compare with Tenenbaum and Griffiths (2001, p. 631). Correspondingly, the generalisation function can be characterised as a sum over a set of conditional probabilities associated with such hypotheses.

$$pr(y \in C|x \in C) = \sum_{h \in \mathcal{H}} pr(h|e), \tag{6.1}$$

Each probability function is associated with a hypothesis, h, given the evidence, e. The term $h \in \mathcal{H}$ indicates a subset of hypotheses in $\mathcal{H}$. Generalisation probability is normalised, so that values of the function in equation 6.1 lie between 0 and 1. Correspondingly, the sum of all conditional probabilities, $pr(h|e)$ must sum up to 1.[4] With regards to the evidence, e, and in the cases of interest here, these are the relevant examples, x and y, for candidate concepts.

Let me describe equation 6.1 with an example. Imagine that an agent observes x, a fly agraric mushroom, followed by y, also a fly agraric mushroom. A hypothesis about x may take the following form: 'x is an instance of the concept FLY AGRARIC MUSHROOM.' A hypothesis about y may take the form: 'y is an instance of the concept FLY AGRARIC MUSHROOM.' The problem of generalising from x to y is the problem of inferring how plausible it is that y is in a candidate concept, C (e.g., FLY AGRARIC MUSHROOM), given that x is in C. Alternative

[4] This implies that any particular conditional probability function must follow the rule that if, for example, $pr(h|e) = .6$ then the sum of the probabilities of all other hypotheses, h^*, must be .4. In other words, the plausibilities of hypotheses in light of the evidence mutually constrain each other.

candidate concepts in this case are the concept MUSHROOM or the concept THING IN THE UNIVERSE.

T&G are not clear about how to interpret the content of the hypotheses. Here, I assume that $\mathcal{H}$ contains hypotheses of the form $h : x \vee y \in C$. That is, $h \in \mathcal{H}$ represents the set of subsets of hypotheses that say that either x or y, or both, are in the extension of the concept C. T&G are also not clear about the status of the generalisation function: they do not make clear whether the conditional probability on the left in equation 6.1 should be interpreted as the probability of a hypothesis given a body of evidence, of the form $pr(h|e)$, or whether it is a probability of a hypothesis given another hypothesis, of the form $pr(h|h)$, instead. Here, I choose the second option and interpret the left-hand term in equation 6.1 as a probability of a hypothesis given another hypothesis. Correspondingly, generalisation is the problem of inferring how probable it is that the hypothesis 'y is in the concept FLY AGRARIC MUSHROOM', h, is true given that the hypothesis 'x is in the concept FLY AGRARIC MUSHROOM', h^*, is true. Correspondingly, if there are only two hypotheses, h and h^*, then the generalisation probability is equal to a weighted sum of the two conditional probabilities, $pr(h|e)$ and $pr(h^*|e)$. Each of these conditional probabilities is a posterior probability associated with a particular hypothesis. The posterior probability represents how plausible a hypothesis is in light of the given pieces of evidence, e.g., in light of the observations of x and y. In the current example, when there are only two conditional-probability functions, then one conditional probability function is associated with the hypothesis that x is in FLY AGRARIC MUSHROOM given the observation of x and the other conditional probability function is associated with the hypothesis that y is in FLY AGRARIC MUSHROOM and the observation of y.

How can equation 6.1 contribute to solving the problem of generalisation? In T&G's model, equation 6.1 represents the agent's psychological categorisation task as a probabilistic problem. In particular, generalisation is now understood as the problem of computing the correlation of the plausibilities of two hypotheses about what concept any of the two objects, x and y (e.g., two fly agraric mushrooms), respectively, most plausibly belong to. For example, we can specify one hypothesis, h, which says that x is an instance of the concept FLY AGRARIC MUSHROOM and the other hypothesis, h^*, which says that y is an instance of the concept FLY AGRARIC MUSHROOM. From the perspective of T&G's model, the task of generalising behaviour from x to y is the problem of identifying the probability of h^* to be true given that h is true. Following equation 6.1, the probability to generalise is equal to the sum of the probability that h is true and the probability that h^* is true. However, implicit in this formulation is that the given object is already known to be an instance of the concept. For instance, when x is the known object, this means that the agent has already figured out in a previous step that x is in the concept FLY AGRARIC MUSHROOM. That is, in this previous step, the agent has already computed the conditional probability that x is in the concept FLY AGRARIC MUSHROOM in light of the observation of x. On the basis of this previous step, the agent 'knows' that it is the concept FLY AGRARIC MUSHROOM with respect to which the relevant hypotheses, h and

h^* correlate. On this basis, the agent identifies how plausible it is that y is an instance of the concept FLY AGRARIC MUSHROOM, given the observation of y. To determine the actual correlation of these hypotheses, the agent then compares the plausibilities by summing over their individual probabilities.

The point of T&G's formulation of the generalisation task (corresponding to equation 6.1) is that a measure of the likelihood of generalising from x to y (the term on the left) can be understood as the average probability that is assigned to each of these hypotheses, h and h^* (on the right). If the left-hand term in equation 6.1 is defined over a hypothesis space with a discrete structure, so that any hypothesis is a statement about the extension of a concept, then the generalisation function (equation 6.1) should be computed with a method known as 'hypotheses averaging' (Appendix A). This method indicates the average of the probabilities associated with all available hypotheses (e.g., those that place x or y in C), weighted by their posterior probability.[5]

So far the discussion has covered how knowledge is represented and how conditional probabilities are transferred into a generalisation function. The third ingredient of T&G's model is a Bayesian approach to concept learning.

AN APPROACH TO CONCEPT LEARNING

By definition, a Bayesian agent should follow Bayes's Theorem in approaching a task (otherwise they would not be 'Bayesian'). T&G (2001, p. 632) characterise part of the generalisation task with the following version of Bayes' Theorem.

$$pr(h|e) = \frac{pr(h)pr(e|h)}{pr(e)} = \frac{pr(h)pr(e|h)}{\sum_{h' \in \mathcal{H}} pr(e|h')pr(h')} \tag{6.2}$$

Equation 6.2 says that the conditional probability of a hypothesis, h, in light of some piece of evidence, e, is equal to the product of the likelihood of observing the evidence given that the hypothesis is true and the prior probability of the hypothesis, regardless of the evidence, taken relative to the sum of the products of likelihoods and priors for all alternative hypotheses, h', in the hypothesis space.

For example, the posterior of the hypothesis that x is an instance of FLY AGRARIC MUSHROOM, given the observation of x, is the ratio of two terms.

1. $pr(x|x \in C)pr(x \in C)$: the product of the likelihood of observing x given that x is an instance of the concept FLY AGRARIC MUSHROOM and the prior of this hypothesis.

[5] T&G suggest that this example is transferable to the case in which the hypotheses space has a continuous structure. In this case, equation 6.1 corresponds to an integral that ranges over all possible candidate concepts in psychological space. In other work (Poth, 2019), I suggest, on the basis of the account of Decock, Douven, and Sznajder (2016), to specify this integral with the Lebesgue integral and the corresponding measure of the concept is the size of a consequential region, as in Shepard's (1987) initial account.

2. $pr(x|x \in C)pr(x \in C) + pr(x|x \in C')pr(x \in C')$: the sum of the products of likelihoods and priors for hypotheses that say that x is an instance of FLY AGRARIC MUSHROOM and the products of likelihoods and priors of hypotheses with alternative candidate concepts, C'. For instance, the concepts MUSHROOM or THING IN THE UNIVERSE.

The same hypothesis can be evaluated in light of another example, y. The posterior probability of the hypothesis that y is an instance of FLY AGRARIC MUSHROOM given the observation of y, is a ratio of the products of likelihoods and priors associated with this particular hypothesis, relative to the sum of the products of likelihoods and priors associated with all alternative hypotheses (e.g., hypotheses that pair y with MUSHROOM or, alternatively, with THING IN THE UNIVERSE). One way of interpreting the posterior probability of a hypothesis is as a representation of the plausibility of a candidate concept that the evidence, x, or y, could be an instance of.

T&G's hormone-levels example illustrates this with some numbers. Assuming that $x = 60$ and $y = 47$, Bayes' Theorem should be used to infer whether each is a healthy (H) hormone level.

$$pr(60 \in \text{HEALTHY}|60) = \frac{pr(60|60 \in H) \times pr(60 \in H)}{pr(60|60 \in H) \times pr(60 \in H) + pr(60|60 \notin H) \times pr(60 \notin H).}$$

$$pr(47 \in \text{HEALTHY}|47) = \frac{pr(47|47 \in H) \times pr(47 \in H)}{pr(47|47 \in H) \times pr(47 \in H) + pr(47|47 \notin H) \times pr(47 \notin H).}$$

What is the posterior probability that a given patient is healthy given that they have a hormone level of 60 (47)? On a Bayesian account, this question should be answered by assessing how likely it is that the patient displays a hormone level of 60 (47) if it was true that the patient is healthy (the likelihood) and how plausible it is that the patient is healthy before their hormone level has been measured (the prior), relative to the sum of the likelihoods and priors for all available hypotheses. In this case, there are only 2 hypotheses: one says that the patient is healthy. The other says that the patient is not healthy.

Bayes' Theorem is a static rule of inference. It specifies probabilities of a single hypothesis in relation to probabilities of other hypotheses at a single moment in time. In the example, this is the time point at which the hormone level 60 is observed and at which the hypothesis '60 is a healthy hormone level' is tested. How can T&G's model contribute to a theory of concept learning?

On a Bayesian approach to learning, a Bayesian agent should follow Bayes' Rule when updating their old degree of belief in a hypothesis at one point in time to a

new degree of belief in the hypothesis at a later point in time.[6] In the context of T&G's paper, Bayes' Rule can be used to explain concept learning (understood as the revision of a degree of belief in a hypothesis in light of new information). For example, learning from a new observation of a level of 47, the agent should update her belief about what it means to be healthy (more precisely, about what objects belong to the extension of the concept HEALTHY).[7]

According to Bayes' Rule the updating of a prior into a posterior probability is moderated by the likelihood function, which measures the plausibility of observing a piece of evidence given that the hypothesis was true. The likelihood function plays a particularly important part in learning because it measures the degree of confirmation that a hypothesis obtains from the evidence. How much the observation of 60 confirms the hypothesis that 60 is a healthy hormone level is indicated by how well this hypothesis predicts the observation of 60. The probability of the hypothesis that 60 is a healthy hormone level in light of novel evidence, 47, will be greater than the prior probability of the hypothesis that 60 is a healthy hormone level when the hypothesis that 60 is a healthy hormone level makes it likely that 47 is a hormone level as well. The crucial part in learning a category is then the choice of the likelihood function. The particular choice of the likelihood function is typically determined by additional assumptions about the process that has generated the examples, x, and y (the underlying sampling process). In T&G's theory, it is assumed that the likelihood function is the size principle and it is assumed that the examples are sampled explicitly from whatever is the true target concept in the inference. Section 6.4 discusses these assumptions in detail.

6.4. T&G's expansion of Shepard's universal law

The first assumption concerns the way in which the examples have been sampled. The sampling assumptions concern the process that has generated the observations x and y from the concepts FLY AGRARIC MUSHROOM, MUSHROOM, or THING IN THE UNIVERSE. T&G's model uses an assumption that is called 'strong sampling' and Shepard's model uses a 'weak sampling' assumption. The role of these assumptions in these models is to guide an agent's inference in a generalisation task. In this section, I contrast and compare these assumptions. I argue that T&G's replacement of Shepard's weak sampling assumption with their strong

[6] In Bayesian epistemology, a Bayesian agent learns a proposition h when in light of new evidence, e, the probability associated with that hypothesis becomes higher or lower than it would be without the new piece of information. Accordingly, h is learned whenever $pr(h|e) \neq pr(h)$.

[7] In the example, the concept HEALTHY develops in this sense, when the plausibility of the hypothesis that 60 is a healthy hormone level given the new observation, 47, is different from the plausibility of the hypothesis that 60 is a healthy hormone level alone. In section 7.3, I argue on this basis that T&G's approach should be interpreted as a theory of concept development, not of concept learning.

sampling assumption reveals that T&G's model is more informative than Shepard's model. In this sense, T&G expand on Shepard's original approach to ULG. Later sections show that the increased informativeness of T&G's model facilitates its application to domains outside generalisation in perceptual domains.

As an overview of the contrast between these assumptions, when generalising from x to y, Shepard's solution is to find the consequential region that is consistent with the example. This implies that if x and y are both instances of both the concepts FLY AGRARIC MUSHROOM and MUSHROOM, then there will be no distinction between how plausible it is that x and y are in the concept FLY AGRARIC MUSHROOM and that they are in the concept MUSHROOM. In contrast, T&G's solution suggests that even when multiple hypotheses are compatible with the evidence, not all of these hypotheses are equally plausible. For example, even when both x and y are instances of both the concept MUSHROOM and FLY AGRARIC MUSHROOM, a hypothesis that assigns them to the concept FLY AGRARIC MUSHROOM is likely to be differently plausible from the hypothesis that assigns these objects to the concept MUSHROOM.

6.4.1. Strong sampling

In the hormone-levels case, the learner assumes that the example, $x = 60$, has been explicitly sampled at random from the category of healthy hormone levels[8]. T&G call this assumption *strong sampling*.

An intuitive example for strong sampling is this. Imagine you go to the forest with a mushroom expert, who shows you a mushroom that has a red head and white spots. The expert tells you that this is a fly agraric mushroom. If you have never seen a fly agraric mushroom before and follow the maxim of strong sampling, you may assume that your teacher has chosen this observation as a positive example of whatever category a fly agraric mushroom belongs to. In this case, the chosen object is a good example of the kinds of mushrooms that are not edible. Even if both the hypothesis that the observed example is a mushroom and the hypothesis that the observed example is a poisonous mushroom are compatible with the observation, the latter hypothesis seems to be intuitively preferable, and safer with regards to avoiding food poisoning.

A second example is a statistical illustration with marbles from Perfors, Tenenbaum, Griffiths, and Xu (2011, pp. 306-307). In this example, concepts are represented by bags of marbles with different colours. Bag A is small and contains a red and a green marble. Bag B is bigger and contains a red, a green and a yellow marble. Given these proportions of differently-coloured marbles in the bag, it is differently likely to blindly pick out a red marble from bag A than to pick out a red marble from bag B. In particular, the probability of picking a red marble given that one reaches from bag A is .5 (since the proportion of red to

[8]The assumption that HEALTHY HORMONE LEVEL corresponds to a natural kind is yet another idealised assumption, which will not be further discussed when illustrating T&G's theory

any colour in the bag is 1 to 2) and the probability of picking a red marble given that one reaches from bag B is .33 (because the proportion of red to any colour in the bag is 1 to 3). Thus, it is more likely to blindly pick a red marble from the smaller bag than from the larger bag.

To illustrate what 'strong sampling' means, Perfors et al. focus on the case of multiple examples, in which the inference task is this. An experimenter randomly samples a sequence of a red, green, red and green marble (with replacement after any pick), from bag A or bag B. The subject has to guess what bag these observations are random samples of. Perfors et al. argue that, given the sizes of the bags (which are known to the subject—it is just not known which bag the sequence has been taken from), bag A should be a better candidate. Intuitively, this is because it would be very surprising to observe this sequence given that the observations were in fact independently sampled from the bag in which the variation of colours is greater—one would expect also a yellow marble to occur in the sequence.

More generally, in strong sampling, examples are chosen independently of each other but not independently of the concept that they are chosen from. This is what is meant by the examples being sampled explicitly from the true concept—the concept that the example belongs to is fixed. In other words, it is not just any concept that is compatible with the example, but a plausible concept.

Contrast T&G's approach to generalisation with Shepard's approach, which takes on the assumption of *weak sampling* instead. According to weak sampling, x happens to fall inside C by coincidence, so that the occurrence of x is independent of the consequential region, C, that contains x. An intuitive example is the mushroom case from above, but consider the slightly different scenario in which you go to the forest with your expert companion and randomly observe a mushroom with a red head and white spots. Is the mushroom edible or is it not? If you have never seen a fly agraric mushroom and follow the maxim of weak sampling, you should expect that your expert companion is going to point out that this observation is a positive example of the kind EDIBLE to the same degree that you expect her to point out this observation as a negative example, and tell you that the mushroom is not edible.

Shepard describes weak sampling along the following lines.

> In finding a novel stimulus to be consequential, the individual learns only that there is some consequential region that overlaps the point in psychological space corresponding to that stimulus. In accordance with whatever probabilities the individual imputes to nature, a priori, the individual can only assume that nature chose the consequential region at random. (Shepard, 1987, p. 1319)

What Shepard means can be intuitively illustrated with the marble case. In weak sampling, the inference task is this. Recall the marbles example from above (with replacement after any pick). In this case, the subject may not know what size the bags have. All that is known is that both bags contain marbles of both red

and green colours and possibly other colours. It is a default assumption, in this case, that the proportions of colours in each bag are equal. Thus, under weak sampling, the subject should infer that both bags are equally plausible candidates for what source the sequence is sampled from.

More generally, under weak sampling, examples (stimuli) are chosen independently of each other and independently of the concept (consequential region) that they happen to be an instance of. An example could have been drawn from any concept that is compatible with the observation, and any such a concept makes this observation plausible in the same way (i.e., it is not the case that among compatible concepts, some are better candidates than others—they are all good candidates).

What is common to the mentioned examples is, intuitively, that both reflect the assumption that the examples have been sampled from the relevant concept with replacement. Thus, in both cases, the marbles are drawn from the bags independently of each other and the occurrence of a mushroom is independent of the occurrence of another mushroom. The difference is that strong sampling adds a dependence relation between the examples and the candidate concepts.

Before moving on, let me clarify why we should care about the distinction between strong and weak sampling from a Bayesian perspective on PC. We should care because each of these distinct assumptions will offer a better fit to distinct sets of generalisation data. This suggests that in different kinds of generalisation tasks, it is reasonable to assume that people use either of these assumptions but not both when they generalise. In the marble example, strong sampling (reflecting the assumption that the marbles are chosen at random within each bag) is more plausible when the contents and relative sizes of the bags are known. If they are unknown, then weak sampling (reflecting the assumption that the chance of the same colour drawn from each bag will be equally likely) is a better assumption. In the mushroom example, the distinction matters in contexts where the inference should eliminate indeterminacy of multiple candidate concepts that are simultaneously compatible with the available evidence. Even when x and y are both in the concept FLY AGRARIC MUSHROOM and THING IN THE UNIVERSE, it must be possible to distinguish between these concepts. Quine (1960) has made this point earlier. There is an infinite number of concepts that can be inferred from the observation of a native shouting 'Gavagai!' while a rabbit runs past (e.g., RABBIT, WHITE, UNDETACHED RABBIT SLICES, ...). Which of these concepts is the right one will depend on additional factors than merely that the concept includes aspects of the observed rabbit. Strong sampling already eliminates among compatible concepts all those that are broad in their intension, while weak sampling suggests that all compatible concepts are plausible.

These examples illustrate that strong and weak sampling seem to depend on different environmental conditions. In the marble case, these conditions concern the contents and relative sizes of the bags. In the mushrooms and Gavagai cases, the conditions depend on assumptions about the context of an utterance such as 'Gavagai!' or about the structure of natural kinds (e.g., UNDETACHED RABBIT

SLICES is intuitively unnatural). When assuming that differences in the conditions under which the data have been sampled matter for the inference, these different assumptions seem differently plausible and their distinction matters.

A formal contrast between the weak and strong sampling assumption is illustrated by Tenenbaum and Griffiths (2001, p. 633), who consider the respective mathematical expressions of these assumptions side by side. Accordingly:

Definition 6.4.1 (weak sampling)**.**

$$pr(e|h) = 1 \text{ if } e \in h \text{ and } 0 \text{ otherwise.} \tag{6.3}$$

Definition 6.4.2 (strong sampling)**.**

$$pr(e|h) = \frac{1}{|h|} \text{ if } e \in h \text{ and } 0 \text{ otherwise.} \tag{6.4}$$

T&G mean by '$|h|$' to indicate the size of a hypothesis, but I think it is more consistent to interpret this term to mean the size of a concept, C, that is indicated by a hypothesis. For instance, when a hypothesis says that x is an instance of HEALTHY HORMONE LEVEL, then '$|h|$' should actually correspond to a representation of the size of the concept HEALTHY HORMONE LEVEL.

The contrast between these equations is this. Weak sampling (equation 6.3) says that if two hypotheses, h_1 and h_2, are consistent with the data, e, then their associated likelihoods should be the same because $pr(e|h_1) = 1$ and $pr(e|h_2) = 1$. For example, assume that h_1 is the hypothesis, 'x is an instance of the concept HEALTHY HORMONE LEVEL', where that concept spans the interval $[58, 62]$ along the hormone-levels scale. Assume that h_2 is the hypothesis, 'x is in the concept HEALTHY HORMONE LEVEL', where that concept spans the interval $[47, 80]$ along the hormone-levels scale instead. Then, under weak sampling, the likelihoods associated with h_1 and h_2 are the same. In other words, there is no way of distinguishing between the plausibilities of these hypotheses as long as they are consistent with the evidence. In contrast, strong sampling (equation 6.4) permits distinctions amongst multiple consistent hypotheses. These distinctions are based on the size (indicated by the absolute value bars of a hypothesis, $|h|$). If $|h_2| > |h_1|$, then $pr(e|h_1) > pr(e|h_2)$. For instance, if the concept proposed by h_1 spans the interval $[58, 62]$ along the scale of hormone levels, while the concept proposed by h_2 spans the interval $[47, 80]$, then, under strong sampling, the likelihood associated with h_1 will be greater than the likelihood associated with h_2. In contrast to weak sampling, the intuition with strong sampling is that the agent takes into account *how much* e is consistent with h.

With respect to the formal contrast, one difference between the two conceptions deserves further attention, namely that the evidence is less informative under weak sampling than under strong sampling. Weak sampling is an all-or-none-like measure of consistency of the evidence with the hypothesis. Strong sampling is a gradual measure of the hypothesis' relative plausibility in light of the data.

Assuming that x has been chosen explicitly helps agents to narrow down the space of all possible concepts compatible with x to only those that, roughly, make the occurrence of x the least surprising. T&G's (2001, p. 633) overview of the formal definitions of strong and weak sampling illustrates the difference. The difference between the informative value of equations 6.3 and 6.4 can be illustrated with an example from Navarro, Dry, and Lee (2012, p. 190), who compare the two ratios of posterior probabilities.

For weak sampling:

$$\frac{pr(h_1|e)}{pr(h_2|e)} = \frac{pr'(h_1)}{pr'(h_2)} \tag{6.5}$$

For strong sampling:

$$\frac{pr(h_1|e)}{pr(h_2|e)} = \frac{|h_2|}{|h_1|} \times \frac{pr(h_1)}{pr(h_2)} \tag{6.6}$$

In equation 6.5, the ratio of the conditional probabilities of the hypotheses given the evidence, $pr(h_1|e)/pr(h_2|e)$, is equal to the ratio of the unconditional posterior probabilities of these hypotheses, $pr'(h_1)/pr'(h_2)$ (not prior probabilities). This is because in determining these posteriors by following Bayes' Rule (see Glossary), the associated likelihoods will be either 1 or 0, so that updating the prior for each hypothesis will not change the posterior in both cases. Thus, observing e does not change the relative plausibility of the two hypotheses. In contrast, in equation 6.6, the relative influence of the likelihoods in determining the posterior changes with a change in the relative sizes of the hypotheses. The ratio of the sizes, $|h_2|/|h_1|$, functions as a weighting on the ratio of the unconditional probabilities of h_1 and h_2. Hypotheses are weighted by their relative sizes and observing e makes a difference to the relative plausibilities of h_1 and h_2 in 6.6 but not in 6.5. Therefore, strong sampling allows the evidence to be more informative for evaluating different hypotheses.

The contrast between weak and strong sampling suggests that choosing between Shepard's and T&G's models reduces to choosing one of these types of sampling assumptions. Which one is more helpful for solving the problem of generalisation (section 6.2.1)? I think that this depends on the context of the problem. The relevant question for evaluating T&G's expansion of Shepard's approach to ULG is, thus, in which context strong sampling is a better method to approach the generalisation task. My claim is motivated by the following thoughts.

In the literature, the two assumptions are sometimes distinguished as a 'learner' and a 'teacher' condition of a generalisation task (M. Frank, Goodman, Lai, & Tenenbaum, 2009; Xu & Tenenbaum, 2007, p. 289). Accordingly, weak sampling is an appropriate assumption when the task does not involve a teacher, but an agent has to infer the concept from an observation alone. Strong sampling is appropriate when a teacher is explicitly teaching the concept, and chooses examples with respect to the goal of communicating the right concept. Navarro et al.

(2012) also propose an intermediate version of weak and strong sampling, which suggests that these two options are not the only possible principles guiding the inference.

One take on this difference is that the two assumptions make different commitments to what psychological processes are involved in the inference task. Accordingly, strong sampling relies on causal reasoning, which is the answer to the generalisation problem when an intention about a concept has caused the observation. For example, observing a fly agraric mushroom is caused by the expert's intention to teach what a poisonous mushroom looks like. In contrast, weak sampling relies on logical reasoning, and does not assume that there is a causal or intentional relationship between the true candidate concept and the observed example. To interpret this in light of the quote from Shepard (section 6.2.1), weak sampling is the answer to the generalisation problem when nature does not intentionally guide observations on the basis of the kinds they belong to. For example, it is a logical possibility that the observed mushroom in the forest is poisonous, but it is also logically possible that the mushroom is edible. When the observed mushroom belongs in fact to the kinds of poisonous mushrooms, this relationship is only a correlation.

A possible alternative to this take is that the different methods are differently optimal solutions for different contexts of the generalisation task, and also depend on the kind of organism that solves the generalisation task (e.g., humans as opposed to pigeons). Correspondingly, the role of these assumptions is to constrain the generalisation function so that it will be optimised relative to the agents' needs and their environments. In these regards, strong sampling seems more plausible in cases where a teacher tries to explicitly teach a concept.

If strong sampling is an expansion of Shepard's approach to ULG, this does not mean that in every case of generalisation, either of these assumptions is better. Thus, the contribution of T&G's model to Shepard's solution is that not only weak sampling, but strong sampling as well, could drive generalisation as a Bayesian inference. However, whichever of these assumptions one chooses as a formal principle in the Bayesian model may commit one to further assumptions about the underlying cognitive process that is involved in the inference, if the formal principle represents aspects of the cognitive-inference process. On T&G's account, the first step to understanding this process is to understand how the relation between the evidence and the content of a hypothesis justifies the assignment of conditional probabilities. For instance, how is it that the relation between the observation of a fly agraric mushroom and the concept FLY AGRARIC MUSHROOM makes the hypothesis that the fly agraric mushroom is an instance of the concept FLY AGRARIC MUSHROOM plausible? T&G's answer is that this function is fulfilled by the size principle, which is a consequence of the strong sampling assumption and which shows that ULG can be derived from this assumption as well.

6.4.2. The size principle

The size principle is a specification of a likelihood associated with each hypothesis in the conditional probability of equation 6.1. For example, the likelihood corresponding to the hypothesis that the object x (e.g., a fly agraric mushroom) is in a concept, C, looks as follows.

$$pr(x|x \in C) \propto \left[\frac{1}{|h_C|}\right]^{|n|}, \tag{6.7}$$

where n corresponds to the number of instances that are given as examples for the concept (cf. Tenenbaum & Griffiths, 2001, p. 633). Equation 6.7 says that the likelihood of observing x given that x was a true random sample of the concept C is proportional to a ratio of the size of the concept that the hypothesis points to, raised to the power of n examples. The same can be done for y. The hypothesis is '$y \in C$' and the corresponding likelihood function is the probability to observe y given that y is an instance of C.

The size principle is a tool for agents to infer the most informative concept when multiple candidate concepts make the evidence likely. The size principle helps to identify, on the basis of the size of a relevant candidate concept, which of these compatible hypotheses should be preferred over others. Roughly, equation 6.7 says that the agent should assign greater probabilities to hypotheses that pair x with a smaller concept. Let me illustrate this with an example.

Suppose x is a fly agraric mushroom, and there are two hypotheses. One hypothesis says that x is in the extension of the concept FLY AGRARIC MUSHROOM. The other hypothesis says that x is in the extension of THING IN THE UNIVERSE. Intuitively, FLY AGRARIC MUSHROOM has a smaller size than THING IN THE UNIVERSE because there are fewer fly agraric mushrooms than things in the universe, whereby the size would correpsond to the concept's extension. T&G suggest to approximate the concept's size with a proxy for the extension, which, on their account, is an unspecified measure of psychological similarity. However, it is unclear on T&G's account, what this psychological proxy is supposed to be. Plausibly, the agent cannot know the extension of a concept—nobody can know how many mushrooms precisely fall under the concept MUSHROOM. Moreover, the extension does not capture the idea that concepts contain possible instances, the ones that have yet not been observed but, upon observation, would fall under the concept. For example, the extension of MUSHROOM does not contain all those (unobserved) mushrooms that could be instances of this concept in the future.

I think that the size of a concept should be understood as a function of the concept's intension. It is widely held that the intension is a quantity of all attributes that the things that could fall under the concept have in common. To make this concrete for a specification of the size of a concept, I build on Shepard's proposal, that concepts are consequential regions in a psychological space that is structured by geometric dimensions, and the relationship between instances of a

concept (i.e., consequential region) is identified with a measure of the geometric distance between their corresponding points in this space. The dimensions in the space represent attributes of (possible) objects. The associated quantity of attributes that are shared is the overall magnitude of the attributes with regards to which the objects that would fall inside a consequential region overlap. In Poth (2019), I have built on Gärdenfors (2000, 2014) Conceptual Spaces approach, in which concepts are also geometric regions, to make this idea more precise. Following this approach, the intension of a concept corresponds to the area that is covered by a geometric region in psychological space. This area can be understood as covering the average similarity among the possible instances of the concept. For instance, the intension of the concept FLY AGRARIC MUSHROOM is the area in psychological space that covers all points that could possibly be observed as instances of the concept FLY AGRARIC MUSHROOM, and can be measured by taking the average of the distances between already existing observations that are known to be in this region. In Poth (2019), I follow Decock et al. (2016)'s application of Carnap's (1980) γ rule, and argue that the size of a concept, that is, the measure of the area covered by a geometric region in conceptual space, can be understood as the Lebesgue-measure over the region. The advantage of this approach to the size of a concept is that similarity (i.e., geometric distance) can be used as a basis for the probabilistic inference process that is used in PC, while, otherwise, the abstract notion of the 'size' of a concept is a rather vague characterisation of the likelihoods in the Bayesian model.

In light of these considerations, and contra T&G's suggestion to use the extension, the reader should henceforth understand the concept's size generally as a function of the concept's intension, although this specification is not part of the Bayesianness of T&G's model of concept learning and rather an interpretation that rests on the assumption of a geometric spaces model of concepts. In this thesis, all subsequent illustrations that appeal to the size of a concept should be understood as expressing aspects that can be associated with the concept's intension under the assumption of a geometric similarity space. For example, the set of mushrooms is on average less similar than the set of fly agraric mushrooms. Thus, following the geometric-spaces conception of concepts, these concepts are different in their sizes because the area covered by the concept MUSHROOM will be larger than the area covered by the concept FLY AGRARIC MUSHROOM.

Given the assumption that instances of a concept occur as systematic random samples in the world (i.e., the assumption of strong sampling) the size principle demands the following. It is quite likely to observe x, a fly agraric mushroom, given that x was a random sample of the category FLY AGRARIC MUSHROOM. But, given the assumption of random sampling, it is very unlikely to observe x if it was randomly sampled from the concept THING IN THE UNIVERSE. In other words, of everything in the universe that you could observe randomly, it is less likely that you observe x than if x was a random observation of things that are fly agraric mushrooms. Therefore, following equation 6.7, the likelihood associated with the hypothesis that x was in FLY AGRARIC MUSHROOM should be higher than the likelihood associated with the observation of x and THING IN

THE UNIVERSE. The size principle can be used to evaluate the plausibility of the hypothesis that y is an instance of the concept as well. Suppose that y is a fly agraric mushroom, and follow the same steps as in the previous example.

The size principle is only a partial answer to how the agent should solve the Bayesian inference task of generalising from x to y. Generalising from x to y means going from $pr(x \in C|x)$ (equation 6.7) to $pr(y \in C|x \in C)$ (equation 6.1). The size principle is a specification of the likelihood that is associated with one of the hypotheses ('$x \in C$' or '$y \in C$', respectively) at a time and the corresponding piece of evidence (x or y, respectively). It is not the generalisation function, which compares this to the plausibility of the hypothesis that y is in C. The size principle determines only partially the plausibilities associated with each of the relevant hypotheses in the inference of generalisation.[9]

In the following, I explicate T&G's view on how this comparison works. My explication is inspired by Shepard's (1987, p. 1319) initial interpretation of generalisation in terms of the relationship between instances and consequential regions.

The comparison comes in multiple layers. Firstly, under the supposition of a geometric space of psychological representations, it holds for any object, x, or y, that this object can be associated with a region, C, in this space. This region represents the candidate concept that x is an instance of. x is known to be an instance of some concept, but it is unknown which concept in particular this is (e.g., FLY AGRARIC MUSHROOM, MUSHROOM, THING IN THE UNIVERSE). For reasons of mathematical elegance and simplicity, Shepard (1987, p. 1319) assumes that the candidate regions centre on x. The second step is an analysis of the size of any candidate region, which is centred on x. Roughly, under the assumption of a geometric space, the size is a measure of the area covered by the region. Thirdly, assuming that x is a random sample of the unknown C, and following T&G's size principle (equation 6.7), it is more plausible that the region centred on x is small, i.e., under the spatial interpretation, that the region covers a rather small area in geometric space.

Doing the same for y, under the assumptions of a geometric space in which y can be located, the availability of regions (i.e., candidate concepts) that centre on y and the size principle, the region that best accounts for the observation y is relatively small rather than large[10]. For example, FLY AGRARIC MUSHROOM is a region around all possible fly agraric mushrooms, centering on y. THING IN THE UNIVERSE is a much larger region and covers, by assumption, about all possible representations in psychological space. Therefore, following the size principle, the probability of observing y given that y is a random instance of the concept FLY AGRARIC MUSHROOM should be higher than the probability of observing y given that y is a random instance of the concept THING IN THE UNIVERSE.

[9] The model is also only implicitly a 'Bayesian' model because priors obtain no explicit attention. It is implicitly assumed that priors are uniform.

[10] It is implicit in the generalisation problem that the region must be larger than x and y. The initial size of C does not depend on the evidence, x or y.

This establishes that a change in the size of the concept leads to a change in the likelihood function that is associated with the observations of x and y, respectively. For example, by replacing the smaller concept, FLY AGRARIC MUSHROOM, with the larger concept, THING IN THE UNIVERSE, the model predicts that the likelihoods, $pr(x|x \in C)$ and $pr(y|y \in C)$ should decrease in both cases.

Taking stock, the previous sections have explained that T&G's model expands on Shepard's earlier model of generalisation with the assumption of strong sampling and the size principle. In the next section, it will be argued that T&G's model offers an approach for rethinking the ULG and psychological similarity.

6.5. From the size principle to generalisation

On the basis of Shepard's geometric account, the key to generalisation is to compare the overlap of the regions centring around x and y, respectively. In this example, x and y are both fly agraric mushrooms. Based on the foregoing analysis, each of them is very likely to be in the region FLY AGRARIC MUSHROOM and very unlikely to be only in the region THING IN THE UNIVERSE. Thus, each of these pieces of evidence is more or less equally likely to be an instance of the concept FLY AGRARIC MUSHROOM and each of these pieces of evidence is more or less equally likely to be an instance of the concept THING IN THE UNIVERSE. In other words, x and y obtain more or less the same likelihoods of being instances of either of these concepts. On Shepard's account, the plausibilities of the concepts in relation to the instances are a measure of the size of the region, which on T&G's account simply corresponds to the inverse of the likelihood. Thus, in intuitive terms and under the interpretation that the generalisation function is a comparison of two hypotheses, equation 6.1 should be interpreted in terms of the relative overlap of the most plausible concepts associated with x and y. If these concepts overlap more, then generalisation probability will increase; x and y will be more probable to be instances of the same concept, rather than of different concepts. If the regions have a small overlap, then it is very improbable that x and y are chosen from the same concept. For a visualisation of the comparison of regions, consider Shepard's (1987, p. 1319) illustration of the overlap in figure 6.1.

Although it is not necessary for the prediction of an increase or decrease of the generalisation probability to make explicit reference to the PS between x and y, it is implicit in this task that the PS between x and y (their distance in geometric space) plays a role in determining the plausibility of x and y to be in the same region. This is because, holding fixed the size of concepts, the closer x and y are in geometric space, the more probable it is that their associated regions overlap.

T&G's framing of the generalisation task suggests that the interpretation of the generalisation function does not require an explicit analysis of PS. From the perspective of the analysis in terms of the size principle, all that is required is a comparison between the likelihoods, which depend on the size of the concept. Their

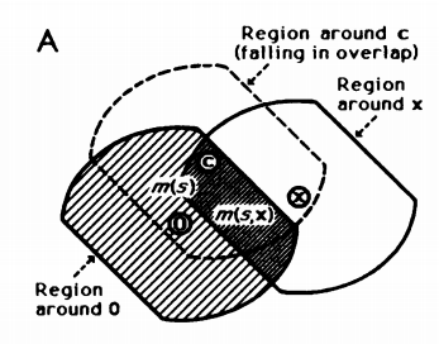


Figure 6.1.: An illustration of the overlap of regions. Region C represents a candidate concept. x and 0 represent the evidence. By assumption, the regions are centred on these examples. $m(s, x)$ measures the overlap of the regions associated with x and 0 and of the same size as C. $m(s)$ represents a measure of the size of the region, C. In this illustration, C functions as a candidate concept for determining the regions around x and 0. From "Toward a universal law of generalization for psychological science," by Shepard, Science, 237(4820), p. 1319. Copyright 1987 by The American Association for the Advancement of Science. Reprinted with permission.

claim is that given the rationale of the size principle, generalisation should follow the exponential gradient (the ULG) as well. Equation 6.7 says that hypotheses that point towards smaller concepts should be assigned a higher probability than hypotheses that point towards larger concepts. This relationship between the size of the concept and the likelihood is exponential and strengthens with more examples (i.e., $x_1, ..., x_n$). The exponential gradient is expected purely on the basis of this formal relation.

Figure 6.2, which I have adapted from Tenenbaum and Griffiths (2001, p. 632), recovers the relationship between the logic of the size principle and generalisation intuitively. The y-axis indicates the probability that y is in C given that x is in C. This represents the model's prediction of how strong the tendency of an agent to generalise behaviour to y in light of x will be. The x-axis represents a hypothetical scale of perceptual representations. x is a point on this scale, representing a perceptual object (e.g., a fly agraric mushroom). The height of a bar indicates the inverse of the size of a concept that is referred to by a hypothesis. Example sizes are indicated on the left of the figure. The length of a bar represents the length of an interval on the x-axis. An interval is a simplified version of Shepard's idea of a consequential region (a concept). The figure illustrates that the probability of generalising from x to y peaks at x and decreases exponentially on both sides.

Intuitively, thicker bars (hypotheses with a smaller size) take up more of the area under the curve associated with the generalisation function, while covering a small interval along the psychological scale. Thinner bars take up less of the area under the 'generalisation curve', while covering a broader interval on the psychological scale. In other words, bars represent different portions of the area

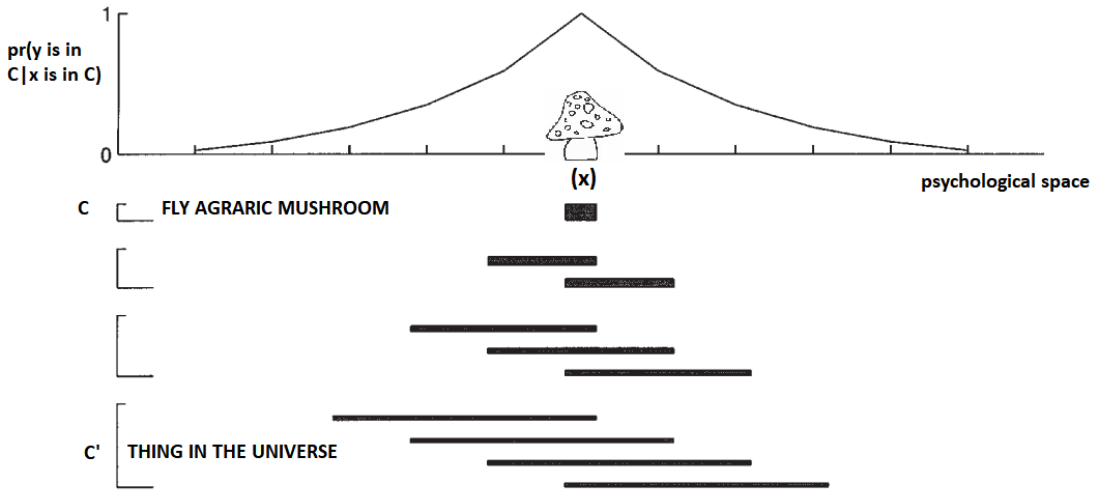


Figure 6.2.: A sketch of how T&G's model accomplishes ULG. The y-axis represents generalisation probability, which exponentially decreases with an increase in the size of a concept, which is depicted as the width of a respective interval covering x. The x-axis represents the values associated with different levels of pigmentations or hormones (see examples in section 6.2). Each interval centres on x and expands equally into both directions along the x-axis. The wider the concept, the further to the left or right of x it will extend, and this is where y-values are small. Adapted from Tenenbaum and Griffiths (2001, p. 631).

in psychological space that is associated with a stronger or weaker generalisation response. The response is stronger for thicker bars and weaker for thinner bars. This expresses the intuition that is associated with the size principle, under the assumption that the inverse of the size of a concept corresponds to the thickness of a bar.

In light of the size principle, the likelihood of observing x, which is a particular point on the scale, given that x was a random sample of one of the candidate intervals along the scale, corresponds to a portion of the area under the generalisation curve. In T&G's model, this area indicates the likelihood. This can be made more specific in light of Shepard's geometric-regions approach, so that it appears that the relevant portion corresponds to the area of the region that is most likely associated with x in psychological space. As can be seen in figure 6.2, most of the area is centred on x. Intervals covering this area more closely obtain higher likelihoods, following the size principle.

A main insight from figure 6.2 is that the more probable it is that x and y belong to the same concept, the more probable is generalisation from one to the other. The probability of an agent to generalise from x to y, on the basis of C, will depend on whether x and y are similarly likely in light of the same concept, instead of different concepts.[11] If both x and y are equally likely to be in C, then

[11]It seems to be a problem for the approach that it is unclear how cases can be solved where the

generalisation will be very probable. If x and y don't overlap in this sense, it will be unlikely that the agent treats them as the same. Following the size principle, the tendency of either x or y to be an instance of the relevant concept is exponential, where the exponent represents the number of examples. Taken together, for both x and y, the generalisation function is therefore almost shaped like a Gaussian with negative exponential curves on both sides. Generalisation probability increases exponentially with an increase in the average (metaphorically: 'overlap') of the likelihoods associated with the concepts centred on x and y (cf. section 6.5). I take it that, on the basis of Shepard's geometric interpretation of PS space, the overlap can be measured as the overlap of the areas covered by the relevant concepts. This suggests to rethink the initial generalisation task: generalising implies measuring the overlap of two likelihoods, as specified by the sizes of the relevant concepts.

However, the examples suggest that this view on generalisation does still implicitly rely on a notion of similarity as a geometric distance between x and y. Intuitively, in figure 6.2, the distance between x and y on the scale determines how much the candidate concepts centering on x and y will overlap. It is because x and y are very close on the spectrum that their associated concepts are most probable to be the same. In other words, given that smaller concepts should be preferred (according to the size principle), the concepts to be compared will be more likely to overlap with each other only if x and y are relatively close in psychological space. Intuitively, longer bars contain more examples than shorter ones because they cover broader intervals in psychological space (the concepts corresponding to longer bars have broader extensions). Thus, the overlap of the likelihoods depends on two aspects. Firstly, following strong sampling, it is less probable for two random points, x and y, located along the x-axis, to fall inside the same interval if the interval centering on each of them is smaller rather than larger. Secondly, the further away (i.e., the less similar) two points are, the less likely they are to be in the same (small) concept.

This interpretation suggests that the generalisation function in figure 6.2 does rely implicitly on the distance (dissimilarity) between x and y. Although this interpretation relies intuitively on the PS between x and y, there is some information that is added by T&G's approach: generalisation becomes a correlation between two likelihood functions that can be identified with the size principle, whereas it had been a measure of the overlap of regions on Shepard's approach. In the next section, I argue that despite its *implicit* reliance on PS, the Bayesian theory of generalisation inspired by T&G's model contributes to a simpler explanation of ULG. I show how this contribution figures in a reverse-engineering approach to PC.

objects are similarly likely in light of two different hypotheses. For example, if it was the case that the probability of observing an apple given that it was an instance of the 'fruit' category was the same as the probability of observing a chair given that this chair was an instance of the category 'kitchen furniture', this would imply, according to T&G's approach, that the apple and the chair should be considered to be similar or obtain the same generalisation response. But they clearly aren't similar and shouldn't receive the same response (e.g, the apple, but not the chair, should be eaten).

6.6. From generalisation to similarity

The derivation of the generalisation function with the size principle is not a mere reformulation of Shepard's results because it adds a new perspective on how the psychological mechanism underlying the generalisation data can be studied. What is new is that the perspective is more abstract: from the perspective of the size principle, no *explicit* definition of PS is needed to understand why the generalisation function should have a negatively exponential shape. At the abstract level of description, reference to the size of concepts and sets of stimuli is sufficient for understanding why generalisation has the shape that it does (figure 6.2). This is because hypotheses pointing to increasingly smaller concepts become exponentially more probable than hypotheses pointing to larger concepts. From this abstract perspective, generalisation is a process of inferring the relationship between x and y depending on the probability that x and y are samples of the same concept. This novel perspective suggests a possible revision of the ULG:

Definition 6.6.1 (The universal law of generalisation revisited)**.** For any hypothesis that pairs one or two stimuli x, or y, with a concept, C, the probability, $pr(y \in C|x \in C)$, to generalise a behaviour from x to y is an exponentially decreasing function of the size of C.

Definition 6.6.1 says that with increasingly bigger concepts, the probability to generalise becomes exponentially smaller. With respect to generalisation, the ULG (definition 3.2) and definition 6.6.1 make similar predictions because, roughly, the meaning of 'size' in T&G's model corresponds to the meaning of 'relative distance' in Shepard's model. However, as indicated above, T&G's model adds emphasis on the relevance of the concept, C, instead of geometric distance. This produces the novel empirical hypothesis that the concept possibly determines the relevant context of the similarity judgement, and that generalisation should change in light of different candidate concepts.[12]

I will now argue that a Bayesian theory of generalisation that is inspired by the conjunction of Shepard's PS model and T&G's model of concept learning exemplifies a strategy to reverse-engineer the cognitive mechanism underlying PC behaviour. Typically, behaviour is recorded in generalisation data and similarity-judgement matrices (cf. chapters 3 and 4) and it is unclear what the physiological processes are that generate the data. Like in Marr's (1982) computational analysis of information-processing systems (chapter 2), the practical implication of a Bayesian approach to PC is that PC can be studied with computational methods, while knowledge about the actual physiological details underlying PC mechanisms is still inaccessible.

[12] One way of elaborating on this idea further is to illustrate that the Bayesian model, in choosing the relevant concept (e.g., an interval along the psychological scale in figure 6.2), determines at least partly what portion of the psychological space is relevant for the inference in the similarity judgement.

In chapter 2, I have illustrated what a computational-level analysis from a Bayesian perspective looks like with the example of vision. Vision is a perceiver's problem to infer from an image on the retina what physical stimulus has caused the image. The perceiver's strategy to solve the problem is Bayesian inference, and the logic is to infer backwards, from features of the image, what the features associated with the stimulus are. The strategy is rational in light of the fact that the true stimulus that has caused the retinal image is not directly accessible, and inferring what stimulus has caused the image will increase the agent's ability to act towards the stimulus in an appropriate way.

The computational-level analysis of PC from a Bayesian perspective proceeds likewise. PC is a perceptual categoriser's problem to infer from a set of examples what concept or category (e.g., FLY AGRARIC MUSHROOM) is associated with the examples. I say 'is associated with the examples' rather than 'has caused the examples' because it cannot be assumed in every case of PC that the relationship between the examples and the concept or category is causal. I have explained in section 6.4.1 that sometimes, it is more sensible to assume that the examples and the category only correlate with each other. The difference is implicit in the additional assumptions about the sampling process—strong sampling assumes a causal relationship, while weak sampling assumes a logical relationship between the examples and a candidate concept.

However, in either of these cases, the perceptual categoriser's strategy to solve the problem is Bayesian inference. The perceptual categoriser must infer the posterior probability of a hypothesis about the 'hidden' category or concept in light of a piece of evidence (i.e., an example of the concept) from the probability of observing the evidence given that the hypothesis is true together with some background knowledge which is encoded in the probability of the hypothesis regardless of the evidence.

The logic of the strategy is to infer backwards, from the similarity among the examples, what the intension of the concept is that is associated with, or has caused, these examples as random samples. Given the distinction between strong and weak sampling, the logic of this strategy can be refined with an optimality principle. This is what the size principle does: when inferring what concept has 'caused' the examples, learners should have a preference for candidate concepts with relatively small intensions. For instance, they should choose DALMATIAN instead of DOG or ANIMAL. On the contrary, in weak sampling, learners should only look out for concepts that are merely 'logically consistent' with the examples. For instance, they may be indifferent in their choice between FLY AGRARIC MUSHROOM or POISONOUS MUSHROOM in light of a few examples that fall under both concepts.

The rationality of this analysis of PC is constrained by additional considerations of what would be the 'optimal' PC behaviour of an ideal perceptual categoriser in a given environment and given their subjective needs. To make this more concrete, I take on Shepard's implicit assumption that concepts fulfil a role for the survival of the perceptual categoriser and assume that PC has adaptive value

(see also preliminaries). In light of the distinction between weak and strong sampling, I distinguish between two principles that are relevant for the rationality of PC. PC behaviour that corresponds to these principles will be optimal and, hence, considered to be rational. The first principle I call 'informativeness'. Following this principle optimises PC performance in cultural environments where the goal of PC is to communicate about perceptual categories. The second principle I call 'survival'. Following this principle optimises PC behaviour in purely 'natural' environments. By assumption, concepts that are informative serve communication, while concepts that are learnable serve the ability to remember. Intuitively, getting the categorisations in 'natural' environments right will increase the categoriser's chances of survival (e.g., by distinguishing edible from poisonous mushrooms). Likewise, in 'cultural' environments, learning to distinguish word-meanings such as 'Dalmatian' from 'dog' increases one's chances of communicating effectively, and also this ability seems to carry adaptive value (Pinker, 2000, 2003). Thus, on the Bayesian approach to PC, PC is rational if and only if it generates behaviour that is optimal by appeal to these additional constraints of survival and communicative success.

The connection between the task of PC and its associated adaptive value can be expressed with the use of a utility function that pairs an expected utility value with each candidate concept, and concepts that are relatively more informative and survival-conducive are assigned to higher expected utility values. A concept learner will choose one candidate concept over another if and only if the expected utility associated with the former exceeds that of the latter concept. The optimal outcome of this function is behaviour produced by a concept that maximises the expected utility with regards to these principles, and the perceptual categoriser should choose this concept because it optimises their chances of inferring the 'correct' (i.e., adaptive) categories.

However, what the 'optimal' combination of values associated with these principles are will depend on the environmental niche. My hypothesis is that Shepard's (1987) weak sampling assumption optimises PC performance with regards to the survival principle, while T&G's (2001) strong sampling assumption optimises PC performance with regards to the informativeness principle. Showing this in detail will require future work, but the general argument is that these two assumptions constrain the Bayesian model of concept learning in ways that make the model produce 'optimal' PC behaviour. This behaviour carries adaptive value in a respective environmental niche.

An example is Xu and Tenenbaum's (2007) Bayesian analysis of word learning, which I discuss in more detail in chapter 7. In Xu and Tenenbaum's experiments, learners have to infer whether, given a sequence of three Dalmatians of the word 'fep', the right concept corresponding to this word is DALMATIAN, DOG or ANIMAL. On Xu and Tenenbaum's Bayesian approach to word learning, learners should follow the size principle and prefer smaller concepts over larger concepts. On my unified approach to PC as a Bayesian inference task, the reason for choosing the smaller concept in this inference task, i.e., DALMATIAN, over the alternative candidate concepts is that this choice maximises the expected utility

associated with the principle of informativeness. In the context of word learning, it is more informative to categorise or label objects more narrowly; for example, in an extreme case, assigning one label to only one object but no other objects will be maximally informative. If the goal of a concept learner is to infer word meanings in such a way that enables them to communicate informatively, then they should favour concepts with relatively small intensions. Favouring concepts with relatively small intensions in the inference of word meaning will help the agent to be informative when using the word to refer to instances of this concept (M. Frank et al., 2009). Therefore, the expected utility should be maximized for concepts that have smaller intensions. In the Bayesian model, the use of the size principle and the assumption that examples are explicit random samples of the true category is justified by the assumption that PC is optimised with respect to communication and word learning, where being informative is most adaptive.

It has been argued elsewhere that in contexts of learning alone, concepts with broader intensions, such as basic-level concepts (Rosch, 1975) produce optimal categorisations. This hypothesis has been extensively studied in the context of language use and the simultaneous elimination of communicative pressures (Kirby, Tamariz, Cornish, & Smith, 2015).

6.7. Conclusion

In summary, this chapter has argued that PC is a Bayesian-inference task that can be analysed at the computational level of explanation (Marr, 1982). I have illustrated this argument with the model of T&G's (2001) *size principle*, which specifies the likelihood function in the inference problem (section 6.4.2). Roughly, the size principle says that given an instance of an unknown concept, generalisation should be stronger for objects that are in the same concept, rather than in different concepts. I have analysed the assumptions of T&G's Bayesian analysis of the generalisation task with an emphasis on the contrast between weak and strong sampling (section 6.4.1). I have argued that there is no unique solution to the generalisation task; strong and weak sampling are aspects of different kinds of generic solutions to the generalisation task given different environmental conditions. On this basis, one interpretation of the additional contribution of T&G's model is that the size principle helps to solve the indeterminacy problem in word-learning tasks (selecting some word meanings over others), while in other cases, weak sampling (equation 6.3) is relatively more plausible (e.g., when robins learn to eat).

I have compared T&G's model to Shepard's model, and argued that the size principle offers a novel, more general, perspective on the ULG (section 6.6). Although the prediction of the generalisation gradient is already available in Shepard's model, it is simpler when derived with the size principle in that it does not explicitly rely on a geometric PS-spaces model. T&G's model derives ULG on the basis of the abstract relationship between likelihoods and sizes of concepts (section 6.5). Generalisation is exponentially decreasing with an increase in the

overlap of likelihoods associated with small concepts, where the size of a concept is assessed intuitively. For example, THING IN THE UNIVERSE has a broader extension than FLY AGRARIC MUSHROOM, therefore, the latter has a smaller size. Thus, the approach is simple because it refrains from specifying the content of the concept precisely.

Finally, I have argued that jointly, Shepard's model of generalisation and T&G's model of concept learning are justifiably part of a reverse-engineering strategy to PC (section 6.6). Together, the combination of these models illustrates an optimal procedure to use generalisation data and reverse infer from that data the underlying cognitive processes that have generated the data. This approach fits into the reverse-inference scheme from chapter 2.

Argument structure of the Bayesian model of PC:

(A) When psychological process, CL, is recruited by a Bayesian-inference task, an exponential gradient of generalisation, E, is likely to be found.

(B) In Bayesian-inference task T, E was found.

(C) Hence, psychological process, CL, was recruited by Bayesian-inference task T.

Where 'CL' refers to a process of concept learning that is driven by the size principle.

To be clear, as part of a reverse-engineering strategy, the Bayesian theory of generalisation should not be seen as a replacement of Shepard's earlier approach to generalisation and PS. At the abstract level of description, the notion of the size in the size principle (equation 6.7) is still unspecified. To make it precise, as I have illustrated on the basis of Shepard's earlier work in section 6.5, a PS space is needed. What is more, to relate the relevant concepts in the hypotheses (e.g., FLY AGRARIC MUSHROOM) to the evidence for the likelihood function, one must still assess some way of identifying whether a concept overlaps, or centres on, the relevant body of evidence (e.g., x or y). The need for defining what 'centrality' and 'overlap' mean calls for an additional theory of how these hypotheses and the relevant concepts are individuated with respect to the sets of stimuli, e.g., Shepard's PS theory. Instead of a replacement, the abstract perspective offered by the size principle should be seen as a contribution to Shepard's geometric PS model. The contribution consists in offering a simpler interpretation of the generalisation data, which is not bound to a geometric-spaces model of concepts.

In sum, the previous discussion suggests that a Bayesian approach to PC systematically combines a theory of PS and generalisation. In the next chapter, I critically evaluate the Bayesian approach to PC.

7. Possible limitations of the Bayesian approach

7.1. Introduction

This chapter evaluates the Bayesian approach to PC presented in chapter 6 with regards to the Shepard-Tversky debate. I argue that the Bayesian approach to PC is positioned at the computational level of explanation. At this level, it can generalise predictions associated with the previous theories of PS but it cannot replace these theories at the representational and algorithmic levels.

To support this argument, I outline four possible limitations of T&G's (2001) model of concept learning from chapter 6. The first limitation is that T&G's strong sampling assumption is too strong in many tasks. This principally limits the size principle to an explanation of word learning (section 7.2). The second limitation concerns the status of T&G's Bayesian model as a theory of concept learning. I suggest that the theory is at best an incomplete theory of concept learning that could be understood as a theory of concept development because it lacks an account of how concepts are acquired in the first place (section 7.3). The third limitation concerns the status of T&G's Bayesian model as a theory of PS and its restriction to the computational level of explanation. It appears that the Bayesian model is limited to a theory of generalisation and does not provide a theory of PS in the sense that Shepard had originally proposed (section 7.4). A fourth limitation is that the prior probabilities are still unspecified in T&G's model (section 7.5). In what follows, I explain each limitation in turn.

7.2. Plausibility of the strong sampling assumption

The first limitation is that T&G's strong sampling assumption, which motivates the size principle, is not genuinely plausible for inference-tasks outside the domain of word learning. Correspondingly, a possible worry for the Bayesian theory of generalisation is that the size principle is restricted to an explanation of generalisation of words or labels, but not of perceptual categories in organisms without linguistic capabilities (e.g., pigeons).

To recapitulate, strong sampling reflects the assumption that perceived objects are explicit random samples of a concept. In section 6.4.1, I have explained that this assumption implies that an observed example of a concept depends causally

on the concept. I have contrasted this assumption with weak sampling, under which the observed instance is independent of the corresponding concept instead.

A case can be made in favour of the strong sampling assumption in the context of communication (see also M. Frank et al., 2009). In their 2007 paper, Xu and Tenenbaum use the strong sampling assumption to explain a Bayesian word-learning task. In their example, $\mathcal{H}$ is a hypothesis space with hypotheses that each pair a candidate concept, C, with a given label, L. A hypothesis can be represented in the form $h_{\langle C,L\rangle}$. For reasons of simplicity, and because the label is always the same in a given word-learning task, L is not explicitly mentioned further.[1] The following are a few examples for such hypotheses.

h_{Dal}: a hypothesis that pairs the concept DALMATIAN with the label 'fep'.

h_D: a hypothesis that pairs the concept DOG with the label 'fep'.

h_A: a hypothesis that pairs the concept ANIMAL with the label 'fep'.

In Xu and Tenenbaum's experiments, word learners are presented with a labelled example and have to indicate for any novel object whether this object should be called with the same label. Xu and Tenenbaum assume that a learner decides this on the basis of the example's perceptual features, a set of available candidate concepts (e.g., DALMATIAN, DOG or ANIMAL) and Bayes' Theorem. Given the evidence (i.e., an object paired with a label), a learner has to compute the probabilities associated with the relevant hypotheses, where each hypothesis pairs the given label with a candidate concept. In the example, the learner has to compute $pr(h_{Dal}|e), pr(h_D|e)$ and $pr(h_A|e)$.

When explaining how learners infer concept-label pairs, Xu and Tenenbaum concentrate on the likelihoods and for this, they use the size principle. They ask: Given a Dalmatian called 'fep', how should a learner assign the probabilities to each of these hypotheses? Following the size principle, if the given evidence is a Dalmatian that is called 'fep', learners should assign the highest probability value to h_{Dal}, and should disregard h_D and h_A. Why? Because the size principle says that given an example, a hypothesis that pairs the label with a smaller concept should be preferred over a hypothesis that pairs the label with a larger concept.

In their paper, Xu & Tenenbaum use the size principle to provide a computational level explanation for their experimental observation that subjects prefer to restrict their generalisations of labels to the smallest possible category level despite the fact that the given evidence is compatible with all hypotheses. For example, they find that, when shown a Dalmatian called 'fep' and subsequently given an array with different kinds of animals to choose from, subjects prefer to generalise the

[1] The label is, however, important in Xu and Tenenbaum's (2007) experiments, where it provides an objective measure of generalisation. (The experimenter uses the label to test whether the learner assigns the same or a different behavioural or labelling response to the training and test objects.)

word 'fep' to all Dalmatians but not all dogs or all animals in the array[2]. Xu & Tenenbaum argue that this result is to be expected under the assumption of strong sampling in a Bayesian model of concept learning. The model would predict that $pr(h_{Dal}) > pr(h_D) > pr(h_A)$ whenever size(DALMATIAN) < size(DOG) < size(ANIMAL). On this basis, Xu and Tenenbaum argue that learners make a rational choice when they generalise 'fep' all and only to Dalmatians in the array, instead of to other animals or dogs.

Xu and Tenenbaum support this argument with a variation in the example. One variation is the case in which learners observe 3 Dalmatians in subsequent order as examples for the meaning of 'fep'. The variation intensifies the intuition that strong sampling is plausible when the task is to eliminate options of the word's meaning. Under the assumption that the Dalmatians are chosen explicitly from the true concept, it would be highly surprising to observe 3 Dalmatians if these were in fact random samples of the category of dogs or the category of animals. Learners think that this observation would be highly surprising under the alternative hypotheses h_D and h_A. Following Xu and Tenenbaum, choosing the hypothesis that pairs 'fep' with DALMATIAN is rational because it helps learners to eliminate the alternative, evidence-consistent, hypotheses efficiently. It is efficient for learners to infer that 'fep' is most plausibly paired with DALMATIAN because only a few examples are necessary to draw this connection.

In this example, the justification of strong sampling rests on additional background assumptions about how words are communicated. In a word learning task, it seems plausible to assume that the teacher knows the true concept and chooses the example (e.g., the Dalmatian) explicitly as an instance thereof. When teaching a word, the teacher presumably intends to be informative, and the learner can assume that the teacher would not have chosen to present three Dalmatians to illustrate the meaning of 'fep', if they had intended to teach the concept DOG, or ANIMAL instead. Assuming that teachers want to be informative in conveying word-meanings justifies the explicit connection between the example and the concept. Xu and Tenenbaum's claim that the size principle enables learners to learn concepts efficiently, from only a few examples, makes the preference of smaller concepts a description of behaviour that is optimal for word-learning.

Nevertheless, T&G's explanation of concept learning seems unjustified for cases outside the domain of word-learning, when there is no agent who could have sampled the example from the true concept. The mushroom example, which contrasts with Xu and Tenenbaum's 'fep' example, illustrates this. If you have never seen a fly agraric mushroom and see one, then this observation, x, should be equally likely to be of the kind EDIBLE as of the kind EDIBLE OR POISONOUS. To someone who lacks the additional knowledge that fly agraric mushrooms are poisonous mushrooms, prior to the observation of x, knowing that x looks like a fly agraric mushroom contributes nothing to eliminating between the competing hypotheses. Intuitively, if there is no reason to prefer either of these hypotheses,

[2]This result is difficult to explain in light of the observation that young children often overgeneralise labels (Bloom, 2002, chapter 1). For a treatment of this issue in the context of the size principle, see Poth and Broessel (2020).

there is no intuition about which of these hypotheses makes the observation less surprising. Correspondingly, there is no reason to believe that it should be more surprising to observe 3 mushrooms that look very similar to x under the hypothesis that these are all edible than under the hypothesis that the 3 mushrooms are edible or poisonous.

In this example, the task is to infer a category outside the context of word-learning and it is difficult to motivate that this context involves communicative intentions. It even seems counter-intuitive to follow the size principle in the example and prefer the smaller concept over the larger concept. Outside the context of a word-learning task, it seems to be more rational, in the sense of adaptively successful, to assume that x, which could be in the concept EDIBLE and also in the concept EDIBLE OR POISONOUS, is an instance of the concept EDIBLE OR POISONOUS. There is no reason to assume that x or any of the other mushrooms is an explicit sample of either of the relevant categories. In this case, when inferring whether x is edible or possibly poisonous, the reason for preferring the concept EDIBLE OR POISONOUS seems to conflict with the principle of choosing the concept with the smallest size (as EDIBLE OR POISONOUS is intuitively larger than either EDIBLE or POISONOUS) and seems to be bound to factors such as adaptive success instead. This suggests that, without further justification, the size principle, which holds only under strong sampling, seems to be limited to a specification of likelihoods in Bayesian word-learning tasks and not in (all) tasks in which one might care about matters of adaptive success on an evolutionary time-scale. Thus, the principle seems to generate behaviour that is 'optimal' only in cases in which the task of perceptual categorisation coincides with a task of word learning.

7.3. A theory of concept development, not acquisition

The second limitation is that T&G present their theory as a theory of concept learning. However, their theory seems to be limited to an account of concept development.

Theories of concept learning typically ask questions about the initial acquisition of concepts from novel experience. The ordinary view is that, when learning a particular concept, one does not have it already. Theories of concept acquisition divide into empiricists and nativists. Empiricists say that experience is necessary for all concept acquisition (e.g. Locke, 1805; Piaget, 1976). Any concept's origin is in experience. Nativists argue that at least some primitive concepts are necessary for experience and for concept acquisition (Carey, 2009; Chomsky, 1986; Fodor, 1975, 2008; Laurence & Margolis, 2002). Moderate nativists and empiricists agree that experience and background knowledge are both sufficient for concept acquisition.

A theory of concept development targets conceptual change in response to new experiences. Theories of concept development target conceptual change, for which

background knowledge is necessary. Such a theory assumes a base of already existing concepts and explains how these concepts change either as a function of experience or as a function of their relations to other concepts, or both. A theory of development may or may not consider experience to be necessary for concept acquisition.

If a theory is a theory of concept learning only if it includes both a theory of concept acquisition and a theory of concept development, then T&G's Bayesian theory does not constitute an ordinary theory of concept learning because it seems to have difficulty to explain the initial acquisition of concepts from experience. This appears to be the case in light of the theory's liability to Fodor's (Fodor, 1975, 2008) puzzle. The puzzle consists of the following premises and conclusion.

(1) Concept learning is a form of hypothesis-formation and -testing. For example, from some observations of ravens that are black, a learner infers: 'all ravens are black.' (2) To represent a hypothesis, one needs to represent its constituting concept (because a hypothesis is a belief and its content is a function of its individual concepts and the way in which they are combined). Thus, testing a hypothesis about a property requires having a concept of that property. For example, if a learner is able to think about the properties BLACK and RAVEN, then that learner has the concepts BLACK and RAVEN (Fodor, 2008, p. 138). The corresponding hypothesis to be tested would be 'the things that are ravens are black'. (1) and (2) taken together are circular. Before testing the hypothesis against the empirical evidence, the concept that should be learned from this experience must already have been available in the first place. The hypothesis 'the things that are ravens are black' cannot be used to learn the constituent concept BLACK. Fodor (ibid.) concludes that concepts are either innate or must be acquired in some other way that does not involve hypothesis testing.

Fodor's conclusion suggests that a hypothesis-testing theory cannot be a theory of concept acquisition. T&G's theory is a Bayesian theory of hypothesis testing that assumes (1) and follows (2). Thereby, it falls short of explaining concept acquisition. In T&G's theory, probabilities can only be assigned to the relevant hypotheses when each hypothesis already takes a candidate concept as its constituent. The hormone-levels example illustrates this. To test the hypothesis that 60 is a healthy hormone level, the doctor needs to know what it means to be healthy. Following premise (2), she must be able to pair the example with a candidate concept, for instance, the interval [55,65]. If Fodor is right[3] and if representing a concept implies having it, then Bayesian inference should not be seen as a theory of concept acquisition, and is possibly incomplete as a theory of concept learning. Therefore, I henceforth use 'learning' in the context of T&G's theory to mean 'development' in the sense of 'learning after initial acquisition'.

Although it is not clear how T&G's (2001) Bayesian model can explain the initial formation of concepts from experience, it has much to say about conceptual change. It explains how novel experiences lead to changes in the relevant beliefs in the inference and in their constituting concepts. In the hormone-levels example,

[3]For criticism, see Margolis and Laurence (2011).

the doctor initially believes that 60 is a random sample from [55,65]. She should change her belief if she takes samples from 3 novel patients who are obviously healthy but have hormone levels of 40, 60 and 77. With this additional evidence, she should be more convinced that [40,80] is the right range of hormone levels that belong to the concept HEALTHY. To explain this change from a Bayesian perspective, it is unnecessary to explain how the doctor acquired her initial representation of the numbers in the interval [55,65]. It is only necessary to explain why, in light of the additional evidence, [40,80] is more plausible than [55,65]. Bayes' Rule provides a guide for this (6.3).

7.4. A theory of generalisation, not psychological similarity

The third possible limitation concerns the explanatory power of T&G's Bayesian model. In particular, T&G's model offers a theory of generalisation but not a theory of PS. One way in which this limitation can be illustrated is based on the distinction between three levels of explanation (cf. chapter 2.2). In the literature on heuristics and biases, the limitation is expressed as a suspicion towards the ability of Bayesian models to explain cognitive processing beyond the level of rational analysis. Brighton and Gigerenzer argue that "a principle objective of the rational analysis of cognition is to narrow down candidate algorithmic level theories by establishing empirically determined performance criteria. If the grand prize in cognitive science is uncovering both why minds do what they do and how they do it, then the productivity and scope of the metaphor would ideally extend to the process level" (Brighton & Gigerenzer, 2008, p. 189). Here, finding out how minds do what they do corresponds to investigations at the algorithmic level or beyond, and, ideally, to a causal explanation of the mental processes that carry out the given computational task (e.g., generalisation as a Bayesian inference)[4].

One might believe that T&G's theory is already operating at the algorithmic level, since the size principle looks like a 'rule' of ranking hypotheses according to the sizes of their concepts. However, there are reasons to believe that this interpretation is too permissive. From the perspective of the Bayesian theory of generalisation, generalisation depends on the probabilities associated with the relevant hypotheses. Regarding ULG, generalisation from x to y depends on how probable it is that x and y are in the same concept, as opposed to different concepts. Bayesian inference derives these predictions on the basis of a criterion of selecting some hypotheses over others, which is the size principle (i.e., hypotheses suggesting smaller concepts should be preferred over hypotheses proposing larger concepts). The tendency to generalise decreases exponentially with an increasing size of a candidate concept. However, the Bayesian theory of generalisation fails to

[4]In a similar vein, Eberhardt and Danks (2011, p. 404) argue that Bayesian models, if not constrained by a rationality assumption, risk to fall back onto methodological behaviourism. On this basis, they associate the unificatory power of Bayesian models in cognitive science with this risk, and call for stronger commitments than unification alone.

specify how hypotheses are individuated with respect to the sizes of the respective concepts. Smaller concepts should be preferred, but what is a smaller concept, as opposed to a larger concept? The size principle theoretically specifies the formal relationship of how some measure of the size of a concept determines the likelihood associated with a hypothesis. But the theory does not say how a hypothesis is individuated; it does not say how the intension or extension of the concept is represented. Therefore, the theory does not include an account of how exactly the rule of ranking hypotheses works. At the level of respresentation and algorithm, a specification of the size of a concept is needed. Thus, the Bayesian model of generalisation is not a theory at this level.

When generalisation is analysed as an abstract problem of Bayesian inference, this does not speak to the mechanisms underlying the corresponding PS judgement behaviour. A comparison with Anderson's (1991) rational model of categorisation illustrates this lack of commitment. Anderson's model has 3 structural components: (a) A level of features, such as [has a beak], [can fly], [has fur] [has four legs]. (b) A level of sets of objects such as [3 robins], [1 penguin], [1 bat], [1 cat, 1 dog, 1 elephant]. (c) A level of category labels: ['bird'], ['mammal']. The model assumes that the categories 'bird' and 'mammal' are organised into subcategories corresponding to the objects (animals) at level (b), which are themselves organised into subcategories of sets of features at level (a). The model also assumes that the resulting classification behaviour corresponds to Bayesian inference over a probability mass function associated with each cluster. For example, since the number of robins is greater than the number of penguins, robin is more probable to show up as a member of the category 'bird'. (Likewise, for 'mammal', bat is unlikely to occur because intuitively, there are fewer bats in the world and so probability will be higher for the set of dog, cat and elephant.)[5] The limited specification of how these clusters of objects could be represented in the brain limits the scope of Anderson's model; Anderson's model offers no explanation of components or activities of the mechanism of categorisation.

Likewise, in the Bayesian model of PC, hypotheses are mathematically abstract variables that do not need to correspond to component parts of a mechanism. The Bayesian analysis of the hormone-levels example in chapter 6 illustrates this. Bayes' Theorem does not state how the doctor actually calculates the relevant probabilities but it says how they should approach the generalisation task. A recapitulation of the likelihood function 6.1, which is key to understanding the task in T&G's model, makes this claim more precise. In chapter 6, I have argued that a notion of PS is already implicit in the likelihood function. My geometric-spaces interpretation of the size principle illustrates this; it shows that the likelihoods associated with each of the two hypotheses, $h_{x \in C}$ and $h_{y \in C}$, depends implicitly on the relative distance of the points, x and y, in geometric space (section 6.5). The reason for why the probability of generalising from x to y becomes greater

[5] This example is simplified and concentrates on an inference over relative frequencies of the objects and their features. Anderson's model counts as Bayesian because it also considers prior probabilities that are associated with the distributions over the sets. For example, in some environments it might be a priori more probable to observe bats than elephants.

with an increase in the correlation between the likelihood functions is the following rationale. If x and y are in the same concept, which should be small, and if that concept is a region in geometric space, then how plausible it is that $h_{x \in C}$ and $h_{y \in C}$ make the observations of x and y, respectively, equally likely depends on the relative distance between x and y in PS space. In other words, on a geometric-spaces interpretation of the relevant concepts, the average of likelihoods in the generalisation function (equation 6.1) depends on the relative geometric distance between x and y.

My analysis illustrates that T&G's proposed explanation of PS with Bayesian inference and the size principle relies on an additional theory of PS processes when going beyond the level of computational analysis to specify the size of a concept. The generalisation function predicts behaviour accurately, and the size principle draws an elegant connection between the likelihoods in the model and a measure of the size of a concept. Nevertheless, it does not specify how the size can be measured, or what kind of representational structure a concept has. To do this, an additional theory of PS (e.g., Shepard's geometric theory) is required to specify the details of the representational structure of the model. Therefore, more justification is needed to explain how Bayesian inference could relate to an identification of a possible mechanism of PS. Nevertheless, the restriction to a theory of generalisation is not fatal for T&G's model. I will show in chapter 9 that the model is still useful to summarise, predict and systematise the observations of the exponential gradient and the directionality effect, which seemed to be in conflict given Shepard's and Tversky's theories of PS.

7.5. Prior probabilities

The fourth limitation of T&G's model of concept learning is that T&G only focus on specifying the likelihoods in Bayes' Theorem but they do not say what could determine the prior probabilities. Prior probabilities could be associated with aspects associated with the size of the concept as well, regardless of whether the concept includes the evidence (x or y). For instance, intuitively and when identifying concepts with Shepard's (1987) consequential regions, the prior probability that some object x and some object y could be instances of C should be greater when C is large and covers many possible points (e.g., possibly both x and y) in PS space. This would mean that hypotheses with larger concepts should a priori be preferred over hypotheses with smaller concepts. However, such a rule would exclude many intuitively plausible hypotheses. One example is the set of hypotheses that is constituted by concepts that include x but not y (or the other way around). Take, for example, the hypothesis h_1, which says that x is an instance of FLY AGRARIC MUSHROOM but that y is not. The alternative hypothesis is h_2 and says that both x and y are instances of the concept MUSHROOM. If prior probabilities should be greater for hypotheses with relatively larger concepts, then, a priori, h_2 should be preferred over h_1 (e.g., because MUSHROOM has an intuitively larger extension than FLY AGRARIC MUSHROOM). Thus, if the prior probabilities

always favour relatively larger concepts, then hypotheses such as h_1 should never be a priori plausible. But this seems too strict. Without additional information, h_1 seems a priori plausible and should not be directly excluded.

What is more, just considering the size of hypotheses cannot account for the intuition that some concepts seem to be a priori more 'natural' than others. For example, in his classical riddle of induction, Goodman (1955) contrasts the predicates 'blue' and 'green' with 'grue' and 'bleen'. To illustrate the riddle, define 'grue' as everything observed before the year 2020 and 'bleen' as everything after the year 2020. Intuitively, all Emeralds that have been observed until now have been found to be green. But at the same time, these observations also fall under the predicate grue. Why have they not been found to be grue? More generally, why is 'green' a more natural predicate to infer than 'grue'? Goodman's answer to the riddle is that 'green' is more projectible, but he did not clearly say what 'projectible' means. The analogy to T&G's case is that hypotheses proposing such unnatural concepts such as GRUE should be intuitively ruled out on an a priori basis, however, there is no criterion in T&G's framework to do this. Notably, the answer cannot have to do with the size of the concepts; 'green' seems to have the same size than 'grue'. Generally, just considering the sizes of the concepts associated with hypotheses in the inference seems to be insufficient to eliminate among them.

There are alternative suggestions for how priors could be specified in the literature on PS spaces, but Tenenbaum and Griffiths (2001) seem to disregard this literature to some extent. One option is mentioned by Shepard (1987, p. 1319), who suggests to include a constraint on the shape of regions (i.e., concepts), when assuming a geometric-spaces framework. According to Shepard, regions in geometric PS should be convex. Roughly, a region is convex if for any two points in the region, any other point falling on a straight line between these two points is, necessarily, also in that region. This option adds another criterion, on the basis of which certain types of regions can be excluded. For instance, it would already exclude star-shaped regions in a PS space. A region that is star-shaped may be very big but still not include some x and y, if x and y would happen to fall in those locations in similarity space that are between the outer ends of the star-shaped region. The criterion of convexity is also preferred by Gärdenfors (2000, pp. 71-72), who identifies convex regions in psychological similarity space with 'natural' properties. Examples for such properties are the concepts GREEN or BLUE as opposed to GRUE or BLEEN. This literature suggests that the size of a concept is plausibly not the only determinant of the prior probabilities, and these are still unknown in T&G's Bayesian model.

The idea that 'natural' concepts are convex regions in similarity space has obtained empirical support from various sources. Sivik and Taft (1994) presented people with a colour patch of the Swedish Natural Color System and a Swedish colour term while asking them to rate along a Likert scale from a to 7 how well the colour of the patch matches the meaning of the term. When modelling subjects' responses in a MDS solution, Sivik and Taft find that none of the data points corresponding to one colour term falls into a region of any other colour

term, thereby concluding that the regions drawn around the data points are convex. A second study that provides empirical evidence for the hypothesis that (some) concepts are convex was conducted by Douven, Wenmackers, Jraissati, and Decock (2016). They focused on the concepts of VASE and BOWL. They first constructed a 3-dimensional city-block space of shape-dimensions, along which they positioned the test stimuli (shapes of vases and bowls that are gradually more or less similar to each other). Participants were shown pictures of shapes and asked which ones they find typical of a vase (or a bowl). Participants were subsequently shown two screens, one with vase shapes and the other with bowl shapes. For each picture, the participant had to click if they found a shape looked typical of a vase or a bowl (for each screen, respectively). Douven et al. subsequently constructed convex hulls of the majority choices of typical vase (or bowl) shapes, and found that nearly all subjects categorised typical vases in the convex hull corresponding to the vase region in 3-dimensional city-block metric space; nearly all subjects categorised typical bowls in the convex hull corresponding to the vase region in the model. These studies provide evidence on categorisation of objects with respect to colour and shape concepts supports the hypothesis that these concepts are convex.[6]

7.6. Conclusion

In summary, this chapter has proposed four possible limitations of the Bayesian theory of generalisation with regards to the Shepard-Tversky debate. The first limitation is that the strong sampling assumption is possibly too strong in tasks outside of the domain of word learning. Secondly, T&G's Bayesian model is better seen as a theory of concept development instead of concept learning. Thirdly, the Bayesian approach inspired by T&G is restricted to a computational level explanation of generalisation data. It is not a theory of PS mechanisms. A fourth limitation is that the prior probabilities are still unspecified in T&G's model. However, the Bayesian approach to PC also contributes to issues in the Shepard-Tversky debate. The first contribution is its instrumental and heuristic value with regards to the Shepard-Tversky debate; it can predict and systematise the findings of the exponential gradient and directionality effects and it can simplify the space of possible representations and principles to instantiate perceptual and linguistic categorisation tasks. The second contribution is that the Bayesian approach is flexible and rivals with the context-sensitivity that had been offered by Tversky's diagnosticity principle.

[6]Although convexity has been proposed as a necessary condition for natural concepts, there is doubt as to whether it is a sufficient criterion of natural concepts, as many non-natural concepts might be convex as well (Douven & Gärdenfors, 2018). For example, Dautriche, Chemla, and Christophe (2016) identify the concept BASEBALL-BAT as a convex concept, and argue that convexity makes concepts more learnable. Douven and Gärdenfors (2018) suggest a set of 'design principles' as additions to the convexity assumption. Whether these principles could play the role of priors in a Bayesian approach to PC is up to further investigation.

What has been learned is, firstly, that the theory of generalisation as a Bayesian inference is a computational level theory that presents a generic solution to the problem of generalisation but stays agnostic about how this solution can be carried out by any particular algorithm of PS. The size principle is only a partial solution to this problem if decoupled from a PS-spaces model and if regarded without deeper investigations of prior knowledge. These aspects limit the Bayesian approach to a computational-level explanation of PC.

In the next two chapters, I argue that the Bayesian approach can contribute to the Shepard-Tversky debate with a unifying theory of the phenomena of the exponential gradient and the effect of directionality. In this way, the Bayesian approach to PC can serve as a tool for predicting and systematising the exponential gradient and the effect of directionality. The next chapter proposes 3 criteria of unification as a test for this argument.

8. Three criteria of unification

8.1. Introduction

In passing, yet without justification, T&G claim that their Bayesian model of concept learning unifies Shepard's geometric and Tversky's feature-matching theories of psychological similarity (PS). They argue that

> [...] when we generalize Shepard's Bayesian analysis from consequential regions in continuous metric spaces to apply to arbitrary consequential subsets, the model comes to look very much like a version of Tversky's set-theoretic models. Making this connection explicit allows us not only to unify the two classically opposing approaches to similarity and generalization, but also to explain some significant aspects of similarity that Tversky's original treatment did not attempt to explain. (Tenenbaum & Griffiths, 2001, p. 336)

More can be done to make T&G's proposal to unify Shepard's and Tversky's models of PS explicit. How is it that their Bayesian model of concept learning unifies Shepard's and Tversky's models? Which criteria of unification does their Bayesian model meet? In what sense is a Bayesian unification of Shepard's and Tversky's theories of PS useful?

In this chapter, I motivate the need for unifying Shepard's and Tversky's theories and outline a set of appropriate criteria for when unification is the case. As a preliminary definition, unification obtains if a theory, T, provides a single set of assumptions to predict two phenomena, p_1 and p_2, while previously, these phenomena had been predicted in light of two distinct sets of assumptions. In the current context, of the distinct sets of assumptions constitutes Shepard's theory of PS as a geometric-distance function. The other set of distinct assumptions constitutes Tversky's theory of PS as a function of matching distinct sets of features.

The conception of unification that I work with in this chapter is based on three different accounts of unification. The first is an account of simplicity or elegance. This account is inspired by Colombo and Hartmann's (2017, henceforth 'C&H') analysis of unification in Bayesian cognitive science. According to this account, Bayesian models of cognition unify because they express complex ideas that are connected to a variety of cognitive phenomena with only a few mathematical equations. The second account is inspired by Kitcher's (1981; 1989) idea that a theory is unifying only if it is of unbounded scope, which roughly means that

the theory predicts many phenomena. I argue that this criterion is intuitively helpful but too imprecise, which motivates the addition of a third criterion of unification. This criterion builds on Myrvold's (2003) Bayesian approach to unification, according to which a theory T unifies two propositions p_1 and p_2 if and only if T reduces the degree of informational irrelevance between p_1 and p_2. This means that, without regards to T, p_1 and p_2 appear to be irrelevant to each other. Unification occurs when, in light of T, p_1 and p_2 become relevant to each other. Taken together, these criteria of unification serve as a test for my argument that a Bayesian theory of generalisation can unify the ULG and the law of directionality.

My incentive to use these three accounts of unification is to gain a better understanding of the aspects that connect and the aspects that compete between Shepard's (1987) and Tversky's (1977) models of PS. Typically, simplicity is motivated by aesthetic or pragmatic considerations. In addition to such considerations, the present account of unification is also motivated by epistemic values associated with theory confirmation and choice. Unification in this context means unification of the phenomena. Therefore, I assess the unifying potential of the Bayesian theory of generalisation not only by its aesthetic and pragmatic standards but also in terms of its ability to relate occurrences of the exponential gradient and directionality effects, while their simultaneous occurrence in similarity-judgement data previously appeared to be puzzling (chapter 5).

The structure of this chapter is as follows. Section 8.1 highlights the differences between the empirical predictions associated with Shepard's and Tversky's models. This section sets the target for the project of a Bayesian unification of the two approaches to PS: to connect the empirical observations of the exponential gradient and directionality in light of a single theoretical background. Section 8.3 discusses the three criteria of unification: elegance, unbounded scope and informational relevance. I conclude that more attention should be paid to the criterion of informational relevance when evaluating to what extent a Bayesian unification of Shepard's and Tversky's models of PS is possible.

8.2. The exponential gradient and directionality effects: two separate phenomena

A major dispute about Shepard's and Tversky's models is whether their empirical predictions about the exponential gradient of generalisation and effects of directionality can be connected. Let us briefly recapitulate how Shepard and Tversky derive these distinct phenomena given their distinct definitions of PS.

As explained in chapter 3, Shepard's (1987) derivation of the exponential gradient transforms ordinal data points (e.g., from Rothkopf's matrix of Morse Code data) into a spatial configuration under the assumption of the metric axioms. The resulting spatial configuration represents pair-wise dissimilarities between

the objects, x and y, as distances between points in a geometric space. The model predicts that the empirical probability of subjects to confuse pairs of objects (i.e., the objective generalisation probability) decreases exponentially with the objects' geometric distance in the spatial configuration.

Assuming that concepts are consequential regions in such a PS space, Shepard (1987) had predicted for any single case, in which an agent has to decide whether she should generalise her behaviour from x to y, that "the conditional probability that $[y]$ is contained in the [candidate] consequential region, given that $[x]$ is, is just the ratio $m(s,[y])/m(s)$ of the (volumetric) measure of the overlap to the measure of a whole such region" (Shepard, 1987, p. 1319). A priori, all possible locations in the consequential region are equally probable to be occupied by a possible instance of the concept, indicating that the agent takes these possible instances to be equally representative of the concept. On this basis, Shepard was able to derive the exponential function in figure 3.1 by integrating over all locations of possible points that the corresponding concept may cover.

Tversky's findings of directionality effects in people's similarity judgements stands in sharp contrast to Shepard's assumption that PS follows the metric axioms, especially the axiom of symmetry (chapter 4). Tversky could account for the observed effects of directionality under the assumptions that PS is a function of a linear contrast between the sets of shared and distinct features that represent the objects in the comparison. Sets of features are discrete entities, unlike continuous dimensions. Tversky had offered two explicit versions of the feature-matching function, but for the current purposes, only the ratio model shall be of interest. To recapitulate, the ratio model says that the PS between two objects, a and b, is the ratio of the set of common features to the sum of their common features, the weighted set of features distinct to a and the weighted set of features distinct to b, where the weights have to be greater or equal to zero. Formally:

$$S(a,b) = \frac{f(A \cap B)}{f(A \cap B) + \alpha f(A - B) + \beta f(B - A)}, \text{ for some } \alpha, \beta \geq 0. \qquad (8.1)$$

Following equation 8.1, if the cardinalities of the sets of distinct features or their associated weights change, then $S(a,b) \neq S(b,a)$ (chapter 4). On this basis, Tversky could accommodate cases such as the observation that people judge the similarity of Tel Aviv to New York to be greater than vice versa. Taken together, in light of Shepard's and Tversky's theories of PS alone, there is no apparent connection or unity between these results of the exponential gradient and the effects of directionality. In the next section, I introduce a scientific conception of unification based on three criteria that I subsequently use to connect these

observations in light of T&G's Bayesian model (chapter 9).

8.3. Three criteria of unification

Current analyses of unification in cognitive science suggest that a Bayesian unification of cognitive phenomena can be achieved in virtue of two theoretical features: elegance (Colombo & Hartmann, 2017) and invariance (Miłkowski, 2016). I claim that Bayesian unification in cognitive science is sometimes justified only if it combines elegance with yet a third criterion: it renders previously separate phenomena or theories informationally relevant to each other (Myrvold, 2003). In particular, this third criterion is needed when the goal of the unification is to combine key insights from previous explanations of these phenomena. Practically, these criteria may not be mutually exclusive. It seems to be helpful to start with a simplification of the theoretical landscape (the first criterion), even when the end goal is to combine previous explanations (the third criterion). For instance, this combination can help to identify key insights of previous explanations of the phenomena and combine previous explanations more efficiently than would otherwise be possible. What is more, if unification is an issue of accounting for a set of phenomena (the second criterion), then the desired combination of the previous explanations (the third criterion) should have an intuitively broad scope in the sense that it should take into account the relevant phenomena to be explained. The task of this section is to explain clearly what the three criteria mean.

8.3.1. Elegance

A common principle of theory choice is to favour simpler theories. There are two broad definitions of simplicity, elegance and parsimony. I focus on the former definition because it is a fairer standard of evaluating the Bayesian theory of generalisation inspired by T&G's model. Below, I explain why.

Simplicity, in the sense of elegance, is a structural property of a theory. Roughly, a theory T's elegance is measured by the syntactic complexity associated with T's auxiliary hypotheses. T is simpler than T' if T contains a fewer number of auxiliary hypotheses, that is, T makes fewer initial assumptions, than T'. An example is given by Baker (2016), who compares the more elegant Keplerian world model and its less elegant Copernican and Ptolemaic predecessors in terms of their theoretical background assumptions.

> The moves from the Ptolemaic model to the Copernican model, and from the Copernican model to the Keplerian model, both involved a reduction in the number of epicycles and free parameters postulated. Since these are both reductions in theoretical apparatus [i.e., in the number and complexity of geometric and astronomical hypotheses], rather than reductions in the number of objects (or kinds of objects) postulated in the world, this amounts in each case to an increase in elegance rather than in parsimony (Baker, 2016, note 15).

In other words, the Keplerian world model is simpler because it reduces the number of hypotheses and laws (e.g., about epicycles and free parameters) that are necessary to account for the observed motion of the planets.

An alternative reading of simplicity is in terms of the number of a theory's ontological commitments. Roughly, a theory is simple if it makes a few ontological commitments. The orthodox view is that a theory's ontological commitment is a demand on the entities that the world has to contain for the theory to be true. This view is typically ascribed to Quine (cf. 1948, pp. 29–33). Correspondingly, if a theory refers to a phenomenon F, then the theory is ontologically committed to F's existence. This means that a theory with a few ontological commitments demands fewer things to exist in the world than a theory with many ontological commitments. This reading of simplicity is commonly described with the term 'parsimony'.

However, the ontological reading of simplicity seems to be relatively inappropriate in some cases of evaluating theories. Such cases occur, in particular, when the evaluation is of mathematical theories. The problem is that the ontological reading of simplicity seems to imply that a mathematical theory that poses an infinite number of entities cannot be parsimonious. To apply an ontological evaluation criterion to a mathematical theory, one would first have to assume some form of Platonism and assume that the postulated entities actually exist. Then, the problem becomes that the mathematical theory would be maximally complex because it would postulate a maximum number of entities. A prominent solution to the problem is to introduce another distinction between an evaluation of the overall *number* of the entities postulated and an evaluation of the *kinds* of entities postulated (e.g., Lewis, 1973, p. 87)[1]. Reducing either makes the respective theory simpler. In Nolan's (1997) words: "Not just total numbers of things, but how many things of *each type* there are is relevant" (Nolan, 1997, p. 340, original emphasis). This suggests a solution to the case of mathematical theories: a mathematical theory that postulates an infinite number of entities may still be relatively more parsimonious than its rivals if it postulates fewer kinds of entities. Thus, also a mathematical theory can possibly be parsimonious in this sense.

But not all mathematical theorists are Platonists. Can these theorists evaluate their mathematical theories with respect to a criterion of simplicity? The notion of 'kinds of entities postulated' seems to rely on the premise that the postulated entities are real because the notion is still an ontological criterion of simplicity. So how would a nominalist evaluate the simplicity of mathematical theories? One way to approach this problem is to assess the mathematical theory in terms of the syntactic structure of the formalisms that the theory postulates. Correspondingly, if a mathematical theory can express an intuitively complex idea with an intuitively simple formalism (e.g., only a single formula), then, intuitively, the mathematical theory is simple. No commitment must be made to the assumption that the formalism describes entities that are ontologically real. I refer to

[1] Originally, Lewis coined the two types of parsimony 'qualitative parsimony' (referring to fewer kinds of postulated entities) and 'quantitative parsimony' (referring to a smaller number of postulated entities) (Lewis, 1973, p. 87).

this sense of syntactic simplicity as 'elegance'. I take the distinction between parsimony and elegance to be important in the current context, in which I evaluate a Bayesian (i.e., mathematical) theory. This is because no commitment will be made that the Bayesian theory describes entities that are ontologically real. Hence, the Bayesian theory under consideration is better evaluated in terms of its elegance, or syntactic simplicity, than with respect to its parsimony.

My idea to evaluate the Bayesian theory in terms of its elegance is inspired by C&H's (2017) previous work on Bayesian unification in cognitive science. According to C&H, "[t]he kind of unification afforded by Bayesian models to cognitive phenomena does not reveal per se the causal structure of a mechanism. The unifying power of the Bayesian approach in cognitive science arises in virtue of the mathematics that it employs: this approach shows how a wide variety of phenomena obey regularities that are captured by few mathematical equations" (Colombo & Hartmann, 2017, p. 462). In this way, Bayesian cognitive scientists can combine observations of many different cognitive phenomena by assigning scientific representations of these phenomena the same types of mathematical properties. In Bayesian decision theory, these properties are typically identified with three abstract ingredients of a Bayesian model. The first ingredient is (a) a hypothesis space, $\mathcal{H}$ (e.g., a set of propositions that are possible to obtain). The second ingredient is (b) a set of prior probabilities, $pr(h)$—prior probability is associated with each hypothesis in the hypothesis space. The third ingredient is (c) a (likelihood) function that relates the space of hypothesis and some pieces of evidence, $pr(e|h)$. Any Bayesian model combines these three ingredients by following Bayes' Theorem, $pr(h|e) = pr(e|h) \times pr(h) / \sum_{h \in \mathcal{H}} pr(e|h) \times pr(h)$ (see Glossary and chapter 6 for an explanation). In Bayesian cognitive science, this method offers a single mathematical description of a variety of cognitive tasks.

Admittedly, C&H do not clearly separate considerations of a criterion of simplicity from other criteria such as unbounded scope[2]. Nevertheless, their argument is useful to explain why a criterion of elegance seems to be better suited than a criterion of parsimony for the purpose of evaluating the simplicity of a Bayesian model. The crucial point in their characterisation of how Bayesian models unify is that the mathematics in Bayesian decision theory is abstract; the mentioned regularities postulated by a Bayesian model need not represent causal regularities in the world. For example, the kind of Bayesian cognitive science that unifies (many phenomena), according to C&H, makes no commitment to the claim that $\mathcal{H}$, $pr(h)$ or $pr(e|h)$ represent ontologically real entities—the mathematics of the model is abstract and cannot be identified with ontological postulates. Thus, I take it that the simplicity of a Bayesian model is better evaluated by analysing the syntactic structure or elegance of the model's formalisms.

In sum, what I have proposed in this section is that a criterion of simplicity should be used to evaluate whether T&G's (2001) Bayesian theory of generalisation can

[2]I think that C&H's description of the abstract mathematics of the model is a description of the model's simplicity. This is because the mathematics of the model can already be characterised without regards to how many types of phenomena (e.g., types of cognitive behaviours) the model describes.

unify Shepard's (1987) and Tversky's (1977) theories of PS. In particular, I have argued that a criterion of elegance is better than a criterion of parsimony to evaluate the simplicity of a Bayesian theory. As a final remark, there is no obvious link between a theory being elegant and it being true. The criterion of elegance (and simplicity more generally) is often appreciated because of its aesthetic value[3]. One of the most intuitive reasons for why the aesthetic character of a theory cannot be identified with a theory's truth is that the former seems to be to some extent subjective while a theory's truth should be a (relatively) objective matter. A unifying theory should not only be aesthetically pleasing but also help us to better understand the phenomena. I take this as a motivation to discuss the criterion of unbounded scope in the next section.

8.3.2. Unbounded scope

Intuitively, a theory has a relatively unbounded scope if the theory can explain or predict many phenomena, or, at least, if it can explain or predict more phenomena than the set of available alternative theories. The main work of cashing out this criterion lies in determining how 'many' or 'more' should be specified.

The argument that a good theory is one of unbounded scope has been discussed rigorously by Kitcher (1981, 1989, pp. 430-519). Kitcher's notion of scope is specific, but for the current purposes it serves to illustrate a traditional conception of unification in terms of unbounded scope. This conception follows the idea that scientific understanding is "the comprehending of a maximum of facts and regularities in terms of a minimum of theoretical concepts and assumptions" (originally Feigl, 1970, 12, cited by Kitcher, 1981, p. 508), and, likewise, that "[u]nderstanding the phenomena is not simply a matter of reducing the 'fundamental incomprehensibilities' but of seeing connections, common patterns, in what initially appeared to be different situations" (Kitcher, 1989, p. 432).

In trying to make these ideas more precise, Kitcher's approach analyses the system of a set of premises and conclusions that a theory consists of and uses to offer an explanation of the phenomena. Roughly, on Kitcher's account, a theory is of unbounded scope (and therefore, on his account, offers a good explanation) if the theory uses only a few argument patterns (on the basis of the available premises) to derive a large number of conclusions.[4] Kitcher's own illustrations are, as he himself admits (Kitcher, 1981, pp. 520–522), complicated. A relatively simple

[3]Sometimes also its pragmatic value—a relatively simpler theory may be preferred over its rivals not because it is closer to the truth but because it is "easier to work with" (Myrvold, 2003, p. 401)

[4]Kitcher takes inspiration for this approach from Friedman (1974), who, according to Kitcher, "argues that a theory of explanation should show how explanation yields understanding, and he suggests that we achieve understanding of the world by reducing the number of facts we have to take as brute." On Kitcher's interpretation, what Friedman could mean by this is that we should "characterize [the explanatory store] as the set of arguments that achieves the best tradeoff between minimizing the number of premises used and maximizing the number of conclusions obtained" Kitcher (1989, p. 431).

example for his approach is the statistical law that almost anyone whose brain is deprived of oxygen for five continuous minutes will sustain brain damage. In application, this law seems to permit quite many conclusions. For example, it can be applied to most animals, so that 'almost anyone' in the argument can be replaced with 'some woman', 'some man', 'some dog', etc.[5] Roughly, from the perspective of Kitcher's account, we can evaluate the power of this explanation by counting the number of its argument patterns and the number of its conclusions. Here, the single argument pattern is the derivation that oxygen deprivation for more than five minutes in an animal typically explains brain damage in that animal, and the number of conclusions is approximally the (high) number of observations of animals to which this conclusion would apply. Thus, the explanation for brain damage as a result of oxygen-deprivation has a broad scope. The corresponding theory of oxygen deprivation in the brain is unifying insofar as the theory can explain more phenomena of oxygen deprivation while using fewer derivations than alternative theories of oxygen deprivation in the brain.

Following this understanding of unbounded scope, the Bayesian models mentioned in the previous section seem to provide good theories. Because of their abstract-mathematical features, Bayesian models can predict a variety of phenomena with a single analytic framework; C&H (2017, p. 454) provide two examples. Firstly, in categorisation, the problem is to infer the probability distribution associated with a set of examples of a category. Secondly, in perception, the problem is to infer the current state of the world given a sensory stimulus-input (e.g., light-reflection or sound-frequency). Both problems can be analysed as Bayesian inference tasks and solved with a generic function that pairs pieces of evidence (e.g., sensory stimuli or category exemplars) and hypotheses (e.g., candidate states of the world or candidate categories), and, in line with the axioms of probability, assigns those pairs a probability distribution. In this way, a Bayesian approach provides a single type of answer (i.e., that the system solves a Bayesian-inference problem) to questions about a variety of phenomena (e.g., perception and categorisation). Thus, intuitively, these theories have a broad scope.

However, it is unclear how their scope can be measured exactly. To recapitulate the previous paragraphs, Kitcher's (1981; 1989) account suggests to measure a theory's scope in terms of the number of phenomena that the theory explains. But this approach to measuring scope is suboptimal for the general purpose of evaluating how good any theory is. This is for two reasons. Firstly, the number of phenomena that a theory can explain is often not critical for how good a theory is. A theory that gives an accurate explanation of a single phenomenon is sometimes as good as a theory that offers explanations of many phenomena. Craver and Kaplan (2018, p. 21) give an example for this while arguing that "[a]n explanatory model that applies only to a single, rare strain of fly is not less explanatory than a model that applies to a ubiquitous strain simply in virtue of the fly's rarity. The model explains fewer tokens, but it explains each of them, we

[5] Some exceptions are freshwater turtles, Arctic ground squirrels, seals and whales and naked mole-rats. All of them have been reported to survive extreme conditions of oxygen deprivation (Larson, Drew, Folkow, Milton, & Park, 2014).

suppose, perfectly well." The model's limited applicability to the phenomenon of a single rare strain of fly makes the corresponding explanation no less powerful than the explanation offered by the more encompassing alternative model. With regards to the distinct phenomena that they explain, both models are adequate in their scope; their explanations seem to be equally powerful with respect to the different contexts in which they apply.

Secondly, it is often the case that the exact number of cases in which a respective theory applies cannot even be identified in the first place. For example, the Big Bang Theory can possibly explain the existence of any galaxy, but we do not know how many possible galaxies there will be whose existence can be explained. There is an uncountably infinite number of phenomena that could be observed in the future that the theory could apply to. If the scope of a theory should tell something about how good the theory is, then the scope of a theory cannot simply be measured in terms of the *number* of phenomena that the theory can predict or explain. Sometimes, other aspects of the theory seem to be more important to evaluate the theory's explanatory scope. Thus, there are reasons to not understand 'unbounded scope' in the literal sense of 'number of phenomena to predict or explain'.[6]

A possible alternative to the number-criterion is to specify the scope of a theory in terms of the theory's *significance*. However, also this alternative is unsatisfying for the current purposes. Let me explain. In trying to illustrate this option, Miłkowski (2016, p. 20) compares the Big Bang Theory with a theory that explains two car accidents in Warsaw. Miłkowski argues that, nominally (i.e., in terms of the number of phenomena that the theory can explain), the Big Bang Theory would have a smaller scope than a theory that explains two car accidents in Warsaw. Against this (unobvious) conclusion about the theories' nominal differences in scope, Miłkowski continues to argue on an intuitive basis that the Big Bang Theory is more powerful because "the Big Bang is of much greater scientific significance" (Miłkowski, 2016, p. 20).

Miłkowski interprets this alternative notion of unbounded scope in the context of current applications of unification strategies, one of which he identifies with Danks' (2014, p. 176) proposal that unification is the application of "some com-

[6]Kitcher seems to be aware of such possible limitations and refines his account by introducing the notion of a *stringent* argument pattern. Accordingly, "unifying power is achieved by generating a large number of accepted sentences as the conclusions of acceptable arguments which instantiate a few, stringent patterns" (Kitcher, 1981, p. 520). Roughly, a stringent argument pattern is one which contains arguments with "some nonlogical expressions and which are fairly similar in terms of logical structure" (Kitcher, 1981, p. 518). On this basis, Kitcher takes up the suggestion that his earlier "conditions on unifying power should be modified, so that, instead of merely counting the number of different patterns in a basis, we pay attention to similarities among these patterns. All the patterns in the basis may contain a common core pattern, that is, each of them may contain some pattern as a subpattern. The unifying power of a basis is obviously increased if some (or all) of the patterns it contains share a common core pattern" (Kitcher, 1981, p. 521). The problem with Kitcher's own alternative to number-counting is that it stays unclear, on his account, how the similarity, or stringency, of an argument pattern, can be assessed.

mon template that is shared by all the individual cognitive models, rather than through shared cognitive elements (representations, processes, or both) across those models". On the basis of Danks' proposal, Miłkowski (2016, p. 22) suggests that a theory's significance is intimately connected to the invariance that the theory offers with respect to the phenomena or sub-theories to which the theory applies. Miłkowski illustrates this connection with an example from Newell's (1994) defence of cognitive architectures as tools for unification within cognitive psychology.

At the heart of Newell's defence is the idea that a unique physical structure can instantiate a variety of information-theoretic tasks. Some examples for these tasks include [1] mental rotation of objects (cf. Cooper & Shepard, 1973), [2] perception of chess positions (Chase & Simon, 1973), [3] linear search on displays (Sternberg, 1980), [4] free recall (e.g. Murdock, 1962), [5] perceiving illusions, e.g. Mueller-Lyer illusion (cf. Berry, 1968), [6] identifying ambiguous figures, e.g. Necker cube and others (cf. Boring, 1943)[7], visual icon (cf. Sperling, 1960). According to Newell, the simultaneous performance of these information-theoretic tasks by a single cognitive architecture offers a single theory about why performance is possible. This theory unifies the psychological sub-theories of these phenomena, which could previously only provide separate answers to how a cognitive system could solve either of these tasks at a time. In light of Newell's example, the theory is invariant in the sense that the same (physical) structure can be used to answer questions about many domains of cognitive processes, and explain why and how the architecture performs all of these tasks simultaneously.

I do not follow Milkowski's approach to unification and unbounded scope for two reasons. Firstly, the invariance of a theory that Miłkowski seems to have in mind is bound to the scope of a theory only if that theory can be positioned at the process-level of explanation. Newell's architecture delivers a theory of the cognitive processes and algorithms that have to be performed to carry out the variety of illustrated tasks (1-7). It is not guaranteed that there is an intimate connection between invariance and the unbounded scope of a computational level theory, such as the one offered by T&G's (2001) Bayesian model of concept learning. However, there is no reason for why a measure of a theory's scope should be bound to a process-level characterisation of the theory. Without further justification for their intimate connection, a separation between Newell's approach to invariance and the general notion of unbounded scope should be maintained. From this perspective, the criterion of unbounded scope can be useful for evaluating a Bayesian theory of generalisation.

The second reason to not follow Milkowski's characterisation of unbounded scope is that there are independent problems with specifying a theory's scope in terms of the theory's significance. One problem is that it is generally not precisely understood what makes a theory significant. This lack of precision is illustrated by Milkowski's intuitive argument that the Big Bang Theory would be nominally of a smaller scope than a theory that explains two car accidents, despite the greater significance of the Big Bang. From Milkowski's illustration, it does not

become clear why the Big Bang Theory is more significant than a theory of two car accidents in Warsaw.

Taken together, my discussion in the last paragraphs reveals that there are currently no satisfying options to precisely define 'unbounded scope'. Nevertheless, I do not consider these objections as fatal to the criterion of unbounded scope; this criterion should not be thrown because the idea that theories of unbounded scope make predictions about a variety of phenomena has some intuitive appeal and historical relevance. In the next section, I consider a third and more precise criterion to evaluate the unificatory status of the Bayesian theory of generalisation inspired by T&G's Bayesian model of concept learning. This is Myrvold's (2003) account of informational relevance.

8.3.3. Informational relevance

On Myrvold's (2003) approach, informational relevance is an epistemic criterion of theory evaluation. What is evaluated is how good a theory is at combining and transferring the information delivered by a set of separate sub-theories or phenomena. Here, I focus on the case in which there are two available sub-theories that are associated with two sets of phenomena. Roughly, a theory, T, unifies two phenomena, p_1 and p_2, if T renders the phenomena informationally relevant to each other.[7] I explain Myrvold's criterion of unification with the notion of probabilistic dependence; T unifies p_1 and p_2 if T reveals them positively probabilistically dependent, whereas p_1 and p_2 appeared to be probabilistically independent before. A special aspect of this criterion is that it relates to issues of theory confirmation and choice. Myrvold also argues that if a theory is better at combining this information than its competitors, then the theory is relatively better confirmed than its competitors. On this basis, the unifying theory might be preferred over the competing sub-theories.

The example of the Copernican theory of planetary motion

This approach to unification can be illustrated with Myrvold's example of the Copernican theory of planetary motion (Myrvold, 2003, pp. 401–406). the Copernican theory, which suggests that the planets orbit the Sun, competes with the Ptolemaic theory, which suggests that the planets orbit the Earth. Both theories can predict equally well the apparent motions of the planets as observed from

[7] The terminology adopted here originates from Myrvold (2003). However, in his original notation, the variable p seems to be used to sometimes refer to either a 'proposition' or a 'phenomenon', a 'hypothesis', a 'sub-theory' or a 'body of evidence'. I use p to represent a proposition that describes a phenomenon. The reasoning behind this is that in Myrvold's Bayesian explication of informational relevance, the degree-of-belief function, $pr(\cdot|\cdot)$, takes propositions (i.e., not phenomena) as its arguments. The informational relevance of p_1 to p_2 in light of T is then always measured with such a function as the amount of information that one proposition delivers to another proposition. I refer to such a proposition sometimes as a sub-theory. I refer to a description of a body of evidence with the variable e.

the Earth. To evaluate which theory is better, Myrvold focuses on the theories' contrasting epistemic values, and asks: how much information does the apparent motion of some planet in the system give about the apparent motion of other planets in the system, given either of these theories? On the Ptolemaic theory, the apparent motion of the planets from the perspective of Earth does not give information about the apparent motion of the other planets and the motions of the planets as observed from Earth appear to be unrelated. In contrast, under the Copernican theory, the apparent motion of the planets can be traced back to the motion of the Earth around the Sun. Under the hypothesis that the Earth itself moves in a circle around the Sun, like the other planets, one can explain the mean deviations of the planets from what would be perfect circles centring on the Sun. Following Myrvold (2003, pp. 405–406), the Copernican theory renders the motion of the Earth around the Sun informationally relevant for the apparent motions of the planets, whereby the theory unifies these phenomena. Myrvold also argues that, in relating the motions of the planets to the motion of the Earth around the Sun, the Copernican theory is mutually supported by the joint set of observations of planetary motions, whereby it obtains overall more evidential support. Therefore, the Copernican theory is better than the Ptolemaic with regards to its unifying power.

Before I explain the details of Myrvold's approach to unification and mutual support, let me mention two remarks that distinguish the criterion of informational relevance from the criteria of simplicity and unbounded scope. Firstly, on Myrvold's approach, unification is not a extra-empirical phenomenon. Typically, the need for unification arises when two theories compete with each other in their explanations of a phenomenon (e.g., planetary motion) and when it is impossible to decide between the two theories on the basis of the evidence alone (e.g., the apparent motions of the planets as perceived from Earth).[8] Often, it is assumed that a choice for one or the other theory must be made on the basis of extra-empirical considerations, such as simplicity, which concerns the aesthetic or pragmatic value of a theory. In contrast, according to the criterion of informational relevance, aspects of the empirical evidence are also relevant for unification: to show that a theory unifies a set of phenomena implies showing that the observed phenomena become in some sense dependent on each other under the unifying theory, whereas they had previously appeared to be independent. As will be argued below, the relevant sense of unification in Myrvold's approach concerns the phenomena's positive probabilistic dependence given the unifying theory.[9]

[8] This is the case in the planetary-motions example but also in the current case of interest: Shepard's and Tversky's theories compete with each other about what similarity is as a psychological phenomenon, and, independently, they both obtain fair amounts of evidence with regards to the data on generalisation and similarity-judgements.

[9] The contrast to simplicity can be illustrated with the example of the Copernican theory. In virtue of the single hypothesis that the orbits of the planets are centred on the Sun, it was possible to predict the deviations of the planets from their mean apparent motions, while these deviations appear to be independent in light of the Ptolemaic hypothesis. The corresponding explanation for planetary motion is simpler (here, more parsimonious) in the case of the Copernican system because the observed daily motion of planets (including that of the Sun) is only attributed to the motion of the Earth instead of other planetary bodies.

Secondly, as an empirically-related criterion, unification becomes relevant for the confirmation of theories. The phenomena play the role of the evidence and the unifying theory and unified sub-theories play the roles of hypotheses or sets of assumptions that may compete with each other. On Myrvold's account, by rendering distinct pieces of evidence or sub-theories informationally relevant to each other, the unifying theory (e.g., the Copernican theory) obtains a higher degree of confirmation by the total evidence (e.g., observations of the motions of the planets in the solar system taken together). This offers a justification to prefer the Copernican theory over the Ptolemaic hypothesis, not on the basis of its simplicity or unbounded scope but by contrasting their relative degrees of confirmation by the evidence taken together (i.e., the total evidence).

In light of these considerations, I explain three key aspects of Myrvold's approach in the next three sections. Firstly, I explain the notion of informational relevance with the notion of probabilistic dependence. Secondly, on the basis of this notion, I explain Myrvold's definition of unificatory power. Finally, I explain how a theory's unificatory power relates to a measure of the theory's evidential support.

Informational relevance as probabilistic dependence

Following Myrvold (2003), a theory unifies two phenomena when the phenomena appear to be informationally irrelevant to each other a priori but appear informationally relevant to each other in light of the unifying theory (Myrvold, 2003, pp. 408–409). Thus, the first step to understanding unification is to understand what it means to render phenomena informationally relevant to each other.

In terms of probabilistic dependence (see Glossary), rendering two propositions, p_1 and p_2, informationally relevant to each other means revealing that they are probabilistically dependent, whereas they had previously appeared to be probabilistically independent (see Glossary). For example, if p_1 describes the apparent motion of Venus and p_2 describes the apparent motion of the Sun, then the descriptions of these phenomena (henceforth only 'the phenomena') are probabilistically independent under the Ptolemaic hypothesis—under this hypothesis, the joint probability associated with these phenomena is no different to the product of the probabilities associated with each phenomenon individually. But under the Copernican hypothesis, these phenomena are modelled as probabilistically dependent. Copernicus assumed that the Earth evolves around the Sun, and that therefore, the motions of the Sun and Earth are related. In light of this assumption, the observed motions of the other planets become dependent on the observer's position relative to the Sun as well. Correspondingly, the probability

But the Copernican hypothesis does more than introducing simplicity: whereas the observed planetary motions appeared to be independent from each other in light of the Ptolemaic previous hypothesis, they suddenly appeared to be dependent on the motion of the Earth in light of the Copernican subsequent hypothesis. In relation to unbounded scope, there seem to be resemblances, as informational relevance implies that the unifying theory will be able to predict the phenomena that it renders relevant to each other. However, Myrvold defines the notion of 'rendering informationally relevant' more clearly.

of observing Venus' motion given the Sun's motion and the Copernican theory should be greater than the probability of observing Venus' motion given the Sun's motion when considered without regards to the Copernican theory.

Let me add two refinements of my interpretation of Myrvold's criterion. Firstly, it is often the case that the way in which a theory renders phenomena informationally relevant to each other is gradual and not absolute, so that my characterisation in the previous paragraph would be too simplistic. A suitable adjustment is to say that rendering p_1 and p_2 to be more informationally relevant to each other means increasing their probabilistic dependence or, conversely, decreasing their probabilistic independence in one's model more than in the competing model. For example, in light of the hypothesis that the Earth moves around the Sun, the apparent motions of Venus and of the Sun are relatively more probabilistically dependent than they had been before, under the competing assumption that the Sun evolves around the Earth.

Secondly, it is useful to introduce a more sensitive distinction between *positive relevance* and *negative relevance*. According to Falk and Bar-Hillel (1983, p. 240), two events, or phenomena, p_1 and p_2 are positively relevant to each other if and only if the probability of p_1 given p_2 is greater than the probability of p_1 alone or when the joint probability of these phenomena is greater than the individual probabilities of each of them taken together. Formally, positive relevance between p_1 and p_2 obtains if and only if $pr(p_1|p_2) > pr(p_1)$ or $pr(p_1, p_2) > pr(p_1)pr(p_2)$. Conversely, negative relevance between p_1 and p_2 obtains if and only if these relationships are reversed, that is, if and only if $pr(p_1|p_2) < pr(p_1)$ or $pr(p_1, p_2) < pr(p_1)pr(p_2)$. For example, the observation of Venus' apparent motion is positively relevant to the observation of the Sun's motion if and only if it is more probable to observe Venus' position, given the Sun's position, than it is to observe each of their positions independently. This is true under the Copernican hypothesis, which traces the joint occurrence of both planets' motions back to a common cause (i.e., the motion of the Earth). This is not the case under the Ptolemaic hypothesis, under which the probability of the Venus' apparent motion and the Sun's apparent motion (at any position) is equal to their individual probabilities taken together. Under the Ptolemaic hypothesis, the phenomena appear to be informationally irrelevant to each other because they are probabilistically independent. Under this refinement, rendering two phenomena positively informationally relevant to each other implies making them positively probabilistically dependent.

Probabilistic dependence and unification

The notion of positive probabilistic relevance explains the link between Myrvold's notion of informational relevance and unification. Accordingly, a theory T unifies two propositions, p_1 and p_2, if T renders p_1 and p_2 positively probabilistically relevant to each other, while they had been either negatively probabilistically relevant or probabilistically irrelevant to each other before. In other words, T unifies

p_1 and p_2 if p_1 and p_2 are probabilistically independent a priori but positively probabilistically dependent in consideration of T.

This explanation of unification pretends that T either unifies or does not unify p_1 and p_2 but Myrvold's notion is gradual. Unifying by rendering informationally relevant (i.e., probabilistically dependent) means reducing the relative informational (i.e., probabilistic) independence of the phenomena more than the competing alternative theories do. Myrvold's (2003, 410) corresponding measure of the unificatory power of a theory T with respect to two phenomena, p_1, and p_2, is captured by the following definition.

Definition 8.3.1 (The unificatory power of a theory considering two types of phenomena[10])**.** The unificatory power, UP, of a theory T, associated with two phenomena, p_1 and p_2, is a measure of the informational relevance, I, between p_1 and p_2 in light of T in contrast to how informationally relevant p_1 and p_2 are to each other alone (i.e., without consideration of T). Formally: $UP(p_1, p_2; T) = I(p_1, p_2|T) - I(p_1, p_2)$.

This definition can be spelled out with the notion of probabilistic dependence introduced earlier. In terms of probabilistic dependence, measuring the degree of unification implies contrasting, on the one hand, how probabilistically dependent p_1 and p_2 are in light of T with, on the other hand, how probabilistically dependent p_1 and p_2 are without regards to T. The relevant targets are conditional probabilities of the form $pr(p_2|p_1, T)$ and $pr(p_1|p_2, T)$ (Myrvold, 2003, p. 401), and the relevant condition of unification is the first in the following list:

(1) p_1 and p_2 are positively probabilistically dependent in light of T if and only if $pr(p_1|p_2, T) > pr(p_1|p_2)$,

(2) p_1 and p_2 are negatively probabilistically dependent in light of T if and only if $pr(p_1|p_2, T) < pr(p_1|p_2)$ and

(3) p_1 and p_2 are probabilistically irrelevant in light of T if and only if $pr(p_1|p_2, T) = pr(p_1|p_2)$,

where each of these relations is symmetric, so that (1) implies that $pr(p_2|p_1, T) > pr(p_2|p_1)$, (2) implies that $pr(p_2|p_1, T) < pr(p_2|p_1)$ and (3) implies that $pr(p_2|p_1, T) = pr(p_2|p_1)$. Unification is the case only if the contrast in definition 8.3.1 is positive, that is, only if p_1 and p_2 are positively probabilistically dependent given T and neither negatively probabilistically dependent nor probabilistically independent. Therefore, when measuring the degree of the unificatory power of T associated with p_1 and p_2, I focus only on condition (1).

The Copernican theory can be used to illustrate this. Let p_1 represent the apparent motion of Venus and p_2 the apparent motion of the Sun. T represents the

[10]Myrvold (2003, p. 412) offers an extension towards multiple phenomena but this is not relevant for the current purposes since the relevant Bayesian unification concerns only two phenomena—instances of ULG and instances of the law of directionality.

Copernican hypothesis that the Earth and the other planets evolve around the Sun. Without regards to the Copernican hypothesis, and under assumption of the Ptolemaic hypothesis, the motions of the planets appear to be probabilistically independent (i.e., condition 3 obtains). The probability of observing, for example, the motion of Venus given the Sun's apparent motion and the Ptolemaic hypothesis that the planets evolve around the Sun should be equal to the probability of observing Venus' motion and the Sun's apparent motion without regards to the Ptolemaic hypothesis. On the contrary, the Copernican theory unifies the motions of the planets because it renders them positively probabilistically relevant to each other (i.e., condition 1 obtains). The probability of observing, for example, the motion of Venus given the Sun's apparent motion and the Copernican hypothesis that Earth and Venus evolve around the Sun should be greater than the probability of observing Venus' motion given the Sun's apparent motion without regards to the Copernican hypothesis. In other words, given the combination of the apparent motion of the Sun and the hypothesis that the Earth moves around the Sun, the apparent motion of Venus is more probable than only given the apparent motion of the Sun. Likewise (because the relationship is symmetric), given the combination of the apparent motion of Venus and the hypothesis that the Earth moves around the Sun, the Sun's apparent motion is more probable than given only the apparent motion of Venus.

A previous interpretation of Myrvold's approach comes from Brössel (2015, p. 529), who uses the notion of conditional dependence. Accordingly, UP in definition 8.3.1 is a ratio of how much "$[p_1]$ and $[p_2]$ are probabilistically dependent under the condition T [...] and the probabilistic dependence of $[p_1]$ and $[p_2]$ unconditionally." In terms of conditional dependence, positive probabilistic dependence between p_1 and p_2 is the case when the truth of p_2 makes it more probable that p_1 is true given T and negative probabilistic dependence is the case when the truth of p_2 makes it less probable that p_1 is true given T. Brössel's conclusion is, likewise, that T unifies p_1 and p_2, if T renders p_1 and p_2 positively probabilistically relevant to each other.

Informational relevance and evidential support

As indicated earlier, what distinguishes Myrvold's approach to unification from the previous alternatives is that it makes unification tentatively relevant for issues on theory confirmation (i.e., questions about how much confidence we should place in a theory). In particular, Myrvold claims that "the ability of the theory to provide a unified account of a set of disparate phenomena contribute[s] to the evidential support these phenomena lend to the theory" (Myrvold, 2003, p. 399). The key argument for this claim is that the support delivered by the evidence or phenomena can be conjoined in favour of the confirmation of the unifying theory. Thereby, the unifying theory picks up relatively more support from the overall available evidence than either of the unified sub-theories do. Correspondingly, if the total empirical support is greater with regards to the unifying theory than with regards to either of the sub-theories, then one may infer that the unified

theory is better confirmed than either of the sub-theories. On this basis, one may choose to work with the unifying theory because it is better supported. The following paragraphs explain this proposal in more detail.

The connection between informational relevance and confirmation theory builds on an analogy between Myrvold's (2003, p. 410) original definition of informational relevance and a classical Bayesian measure of confirmation, which is $pr(h|e)/pr(h)$ (Strevens, 2012, p. 45). The analogy can be seen in two steps. Firstly, under the classical Bayesian measure of confirmation, a hypothesis, h, is confirmed (to some degree) by a piece of evidence, e if and only if the prior probability of h conditional on e is greater than the prior unconditional probability of h (i.e., e confirms h if and only if $pr(h|e) > pr(h)$). On Myrvold's approach, the role of the evidence is played by the phenomena, for instance, the apparent motion of Venus can be represented by e_1 and the apparent motion of the Sun can be represented by e_2. Let T represent the hypothesis in question (e.g., the Copernican hypothesis).

Myrvold then assumes that the degree of evidential support for T by e_1 and e_2 taken together can be measured with the classical Bayesian measure of confirmation. Correspondingly, T is confirmed by e_1 and e_2 if and only if T is a priori unconditionally less probable than conditional on e_1 and e_2. Likewise, an alternative theory, p_1, is confirmed by e_1 if and only if p_1 is a priori unconditionally less probable than conditional on e_1. The same holds for the relation between another theory, p_2, and the corresponding piece of evidence, e_2[11].

Myrvold also assumes that the support offered by the independent pieces of evidence, e_1 and e_2, is additive. An additivity condition is common in Bayesian confirmation theory, for example, the Bayesian log likelihood measure of confirmation is additive (Strevens, 2012, p. 45). Myrvold follows this example and assumes that the measure of informational relevance is additive: "when $[e_1]$ and $[e_2]$ are independent items of evidence, the information yielded about $[p_2]$ by the conjunction of $[e_1]$ and $[e_2]$ will be simply the sum of the information yielded by $[e_1]$ and the information yielded by $[e_2]$" (Myrvold, 2003, p. 409).

Secondly, Myrvold identifies the logarithm of the classical Bayesian measure of confirmation with the measure of informational relevance (Myrvold, 2003, p. 411). Correspondingly, the logarithm of $pr(h|e)/pr(h)$ measures how informationally relevant e is to h, and the measure of informational relevance, $I(h, e)$, becomes

[11] My notation is an adaptation from Myrvold. His original formulation of the classical Bayesian measure of confirmation is $Pr(h|e\&b)/Pr(h|b)$, where b represents some theoretical background (Myrvold, 2003, p. 411). Myrvold argues that "The background b need not be the sum total of facts known to an agent at some time, and, in particular, should not include the evidence e being considered" (Myrvold, 2003, p. 401). It is common to assume that the background is already accepted as a rational agent's total knowledge, so that $pr(b) = 1$. I adopt this assumption for matters of simplicity. Therefore, I do not mention b explicitly when explaining Myrvold's account. As far as I can see, my simplification does not change the general implications of Myrvold's proposal to identify a measure of informational relevance or probabilistic dependence with the classical Bayesian measure of confirmation.

"a candidate measure of the degree to which a piece of evidence supports a hypothesis" (Myrvold, 2003, p. 412). In the current case, the outlined analogy between informational relevance and the classical Bayesian measure of confirmation concerns the conjunction of e_1 and e_2 and the theory T, so that the classical Bayesian measure of confirmation can be compared to the informational relevance of the total evidence on T. Formally: $\log pr(T|e_1 \wedge e_2)/pr(T) \approx I(T, e_1 \wedge e_2)$. A third condition seems to be implicit in Myrvold's approach; that p_1 and p_2 compete with each other, that is, when p_1 is confirmed by e_1 but not by e_2, and p_2 is confirmed by e_2 but not by e_1.

If the analogy is suitable (e.g., based on the structural resemblances of the formalisms), and under the conditions that the evidence is additive and sub-theories compete, the measure of informational relevance can be used to assess the relative degree of confirmation of T by the conjunction of e_1 and e_2 as compared with the previous sub-theories, p_1 and p_2, and their competing explanations of the available evidence. In other words, if $I(T, e_1 \wedge e_2)$ is relatively greater than either $I(p_1, e_2)$ or $I(p_2, e_1)$, then we could infer that T is better confirmed by the joint occurrence of e_1 and e_2 than either p_1, by the occurrence of e_2, or p_2, by the occurrence of e_1.

Connecting these two steps makes explicit that Myrvold's measure of UP (definition 8.3.1) reveals something about the relative degree of confirmation of the theory by the conjunction of the pieces of evidence as contrasted with how much these pieces of evidence confirm the competitive sub-theories, respectively. When we assume that the likelihoods associated with p_1, p_2 and T are equal, so that $pr(e_1|T) = pr(e_1|p_1)$, $pr(e_2|T) = pr(e_2|p_2)$ and $pr(e_1|p_1) = pr(e_2|p_2)$, then no decision about which theory to choose can be made on the basis of the available evidence. However, following Myrvold's account, "if [T] unifies the pair $\{e_1, e_2\}$ by making [these observations] informationally relevant to each other, and [neither p_1 not p_2 do], then the likelihood of [T] on the [conjunction of the] evidence e_1 & e_2 is higher than that of [either p_1 or p_2], and consequently T is better supported by e_1 and e_2 than [either p_1 or p_2 are]" (Myrvold, 2003, p. 412).

The link to confirmation indicates a possibility to choose the unifying theory over alternative theories on the basis of the relative support that these competitive theories obtain from the available evidence. The additional evidential support for the unifying theory suggests that the unifying theory is better confirmed. If theory confirmation is a criterion of theory choice, then the unifying theory should be accepted in disfavour of the competing sub-theories based on the total available evidence.

8.4. Conclusion

In the context of the overall thesis, this chapter has served the goal of making explicit how the theory of generalisation as a Bayesian inference problem, inspired by T&G's Bayesian model of concept learning (chapter 6), can unify Shepard's

and Tversky's theories of PS (chapters 3 and 4), by outlining a set of criteria for when unification is the case.

In summary, I have explicated 3 criteria of unification: simplicity, unbounded scope and informational relevance. I have argued that the notion of elegance or syntactic simplicity is particularly useful for evaluating how simple the Bayesian theory of generalisation is. In explaining the popular interpretation of unbounded scope in terms of the number of phenomena that a theory can explain, I have used Kitcher's (1981; 1989) approach to explanation as unification. I have rejected Milkowski's (2016) suggestion to interpret the criterion of unbounded scope in terms of a theory's invariance, which he had based on an illustration of Newell's cognitive architecture. My argument against a specification of unbounded scope in terms of invariance has been that the example from Newell's architecture is bound to the assumption that the relevant theory under inspection is a process-level theory (i.e., a theory that is able to specify the algorithms that produce the relevant cognitive behaviour). If bound to the process-level, the criterion of invariance would be inappropriate for evaluating T&G's Bayesian model of concept learning, which has not crossed the boundaries associated with the computational level of explanation (chapter 7). Thus, for the purpose of evaluating a Bayesian theory of generalisation that builds on T&G's model, the criterion of unbounded scope should be kept separate from a criterion of invariance.

I have argued that the criterion of unbounded scope is yet not well enough understood but it has intuitive appeal. On this ground, I have added a third and more precise criterion to the discussion. This is the criterion of informational relevance, based on Myrvold's (2003) Bayesian approach to unification. I have explicated this criterion with the example of planetary motion and the definition of conditional dependence. I have explained Myrvold's argument that this approach indicates a possibility to choose the unifying theory over the competing sub-theories on the basis of the relative evidential support that these competitive theories, in contrast to the unifying theory, obtain from the available evidence. The key to this argument is that, under certain assumptions, the conjunction of the bodies of evidence that were originally associated with each of the sub-theories delivers additional evidential support for the unifying theory.

Taken together, the combination of all three criteria is relevant for evaluating whether a Bayesian theory of generalisation inspired by T&G's Bayesian model of concept learning can unify Shepard's and Tversky's theories of PS. In the next chapter, I apply these three criteria to the advocated Bayesian theory of generalisation and argue that this theory earns the label of a unified theory.

9. Unifying perceptual categorisation

9.1. Introduction

The task of this chapter is to show that the Bayesian theory of generalisation satisfies the three criteria of unification that were presented in chapter 8. I argue that the theory thereby unifies the phenomena of the exponential gradient of generalisation and the effect of directionality in similarity judgements, while these phenomena had previously appeared to be unrelated under Shepard's and Tversky's models of psychological similarity (PS).

My argument that the Bayesian theory of generalisation advocated in chapter 6 unifies Shepard's and Tversky's competing accounts of PS goes like this. Firstly, I show that the Bayesian theory of generalisation provides an elegant analysis of the cognitive task: generalisation is a Bayesian inference task (section 9.2). Secondly, I show that the theory of generalisation as a Bayesian inference has a broad scope because it can simultaneously predict both, instances of the ULG and instances of the law of directionality (9.3). Thirdly, I show that the Bayesian theory of generalisation makes the occurrence of the exponential gradient appear to be positively probabilistically dependent on the effect of directionality, while these observations had previously appeared to be independent of each other (section 9.4). I conclude that the proposed Bayesian approach is a third alternative to Shepard's and Tversky's proposed explanations of these behaviours.

9.2. Satisfying the first criterion

From the perspective of a Bayesian approach, the problem common to Shepard's and Tversky's approaches is a problem of generalisation, and its solution is Bayesian inference. From this perspective, the Bayesian theory of generalisation inspired by T&G's model is syntactically simpler or more elegant than the disjunction of Shepard's and Tversky's models of PS. This is because the Bayesian theory of generalisation makes fewer initial assumptions about what generalisation is and how it works than either Shepard's or Tversky's models make about what PS is and how it works.

In particular, the disjunction of Shepard's and Tversky's assumptions regarding the phenomena of the exponential gradient and the effects of directionality seems

to be relatively more complex than the set of assumptions that constitute the Bayesian theory of generalisation. This contrast becomes apparent when considering the following conflicting sets of assumptions about aspects of PS.

The structure of representational space:

(a) geometric in Shepard's model, and

(a') set-theoretic in Tversky's model.

The primary entities:

(b) In Shepard's model, these are continuous dimensions with a distance metric.

(b') In Tversky's model, these are discrete sets of features from a decomposable data-base of objects.

The form of the PS function:

(c) In Shepard's model, PS is derived from an exponential function.

(c') In Tversky's model, PS is a linear function of matching sets of features.

The explanatory target:

(d) The agent's task in Shepard's model is to generalise from a to b.

(d') In Tversky's model, the task is to judge how similar a is to b (in one direction) or how similar b is to a (in the other direction).

Together, the set of these two subsets of mutually exclusive or conflicting alternatives, $\{a, b, c, d\}$ and $\{a', b', c', d'\}$, is a relatively complex representation of Shepard's and Tversky's distinct assumptions about how PS should be defined.

In contrast to the disjunction of Shepard's and Tversky's models, the Bayesian theory of generalisation inspired by T&G's model builds on a single set of three assumptions about what is needed to solve a Bayesian generalisation task. These three assumptions are that learners should be treated as (a*) using a hypothesis space, $\mathcal{H}$ (e.g., a set of propositions about possible candidate concepts), (b*) using a set of prior probabilities, $pr(h)$, associated with each hypothesis in the hypothesis space, and (c*) using a relation between hypotheses about candidate concepts to pieces of evidence with the size principle, $pr(e|h) = (1/\text{size}(h_C))^{|n|}$, for scoring hypotheses according to the size of a concept (chapter 6). Learners are treated in this way to describe the function that they compute. However, this description implies no commitment to the claim that this is how a learner actually computes the task in their minds.

In T&G's model of concept learning, each hypothesis in $\mathcal{H}$ is a proposition about a candidate concept, C. For example, two hypotheses, h_F and h_T, express different statements about the membership of x to either of two candidate concepts, F and T.

h_F: x is in the extension of the concept FLY AGRARIC MUSHROOM.

h_T: x is in the extension of the concept THING IN THE UNIVERSE.

These hypotheses are different because the extension of FLY AGRARIC MUSHROOM is (intuitively) smaller than the extension of THING IN THE UNIVERSE. Each hypothesis places a relevant object (e.g., x) in a particular candidate concept (e.g., THING IN THE UNIVERSE). Candidate concepts can overlap (e.g., the concept THING IN THE UNIVERSE entails the concept FLY AGRARIC MUSHROOM). Correspondingly, hypotheses pick out concepts of varying sizes.

The probability function in (b*) is defined over $\mathcal{H}$. On T&G's account, the prior distribution of probabilities associated with these hypotheses is uniform, so that there is initially no distinction between how plausible the concepts posed by h_F and h_T are when x has not yet been observed. (c*) relates each of these hypotheses to the evidence, x, or y, respectively, to determine the likelihoods of these hypotheses.

So far it looks as if T&G's model makes only three basic assumptions but it is fair to characterise their model with a fourth assumption: that the generalisation task can be analysed with Bayes' Theorem to begin with.[1] The assumption is that in the specific case of a Bayesian inference of a generalisation problem, learners ask themselves: what is the probability that y belongs to the concept FLY AGRARIC MUSHROOM (alternatively, to THING IN THE UNIVERSE), given that x does? Following T&G's model, the corresponding Bayesian generalisation function takes the following form.

$$pr(h_{y\in C}|h_{x\in C}), \tag{9.1}$$

where '$h_{y\in C}$' is the hypothesis that the novel stimulus, y, is in a candidate concept, C, and '$h_{x\in C}$' represents the hypothesis that the old stimulus, x, is in C. Expression 9.1 says that generalisation is a task of inferring the conditional probability of the hypothesis that the novel stimulus, y, is in candidate concept, C, given the hypothesis that x is in C. On the basis of the agent's prior acceptance that x is in C, and given some way of comparing her observations of x and y with respect to this concept (e.g., in terms of the relative overlap of the concepts that make these observations most likely, as I have suggested in chapter 6), the agent infers whether y is also in C. Thus, under the assumption that generalisation is a Bayesian inference, the task of generalising from x to y reduces to the task of inferring the extension of the relevant concept that x belongs to and deciding whether y also belongs to the extension of this concept[2]. As argued in chapter 6,

[1] To recapitulate, Bayes' Theorem says that $pr(h|x) = pr(x|h)pr(h)/\sum pr(x|h)pr(h)$. In the general context of some inference task, Bayes' Theorem says that the posterior probability of some hypothesis, h, in light of some body of evidence, x, is equal to the ratio of the product of the likelihood of observing x, given that h is true, to the product of the likelihood and prior associated with all hypotheses in $\mathcal{H}$. (See chapter 6 and Glossary for a detailed exposition.)

[2] The Bayesian model stays largely agnostic about how the extension or intension can be specified. The agent will at most only have access to a proxy for the extension (e.g., it cannot be known how many fly agraric mushrooms there are actually in this world). I have

a partial solution to this task is to generically compute the conditional probabilities associated with the likelihoods of the relevant hypotheses (e.g., h_F and h_T, for x and y respectively), as determined with the size principle.

Taken together, the Bayesian model uses a single set of coherent assumptions, whereas the disjunction of Shepard's and Tversky's models contains two mutually exclusive/conflicting sets of assumptions (i.e., two distinct structures of the representational space, two distinct types of primary entities, two distinct forms of the PS function and two distinct explanatory targets). Together, these are 8 distinct assumptions about aspects of possible processes associated with PS. In contrast, the Bayesian model provides a single set of four assumptions: generalisation is a task of Bayesian inference, which can be analysed with three ingredients (a*, b* and c*). Overall, these are only 4 coherent assumptions about aspects associated with generalisation. This shows that the characterisation of generalisation as a Bayesian inference task is relatively less complex when compared to the disjunction of Shepard's and Tversky's characterisations of PS. Because the Bayesian theory may principally pose an infinite number of hypotheses, we should describe the Bayesian analysis as relatively more elegant than the disjunction of Shepard's and Tversky's descriptions of PS (see discussion in chapter 8).

9.3. Satisfying the second criterion

The second criterion of unification is unbounded scope: the idea that a theory unifies only if it predicts more phenomena than its available alternatives. This idea can be applied to the Bayesian theory: if the theory is of relatively unbounded scope, it should predict more phenomena than either Shepard's or Tversky's theories do. This section shows that the Bayesian model scores well with regards to the criterion of unbounded scope because it predicts that generalisation from x to y has an exponential shape and that generalisation from x to y is likely to be different to generalisation from y to x (i.e., generalisation is likely to be directional). In contrast, these phenomena could only be derived one at a time with either Shepard's or Tversky's earlier theories.

T&G's model makes these predictions with two versions of this Bayesian analysis of the generalisation task (expression 9.1). One version of T&G's analysis of generalisation as a task of Bayesian inference (expression 9.1) accommodates instances of the ULG. I have illustrated this with a mushroom example in chapter 6 (cf. figure 6.2). When comparing the probabilities associated with the hypothesis that a fly agraric mushroom, x, belongs to the concept FLY AGRARIC MUSHROOM, and the hypothesis that a fly agraric mushroom, y, is also an instance of the concept FLY AGRARIC MUSHROOM, it is the relative overlap of the likelihoods associated with these hypotheses that predicts the exponential gradient associated with ULG. In the example from chapter 6, a concept corresponds

suggested in chapter 6 that we should understand concepts in the model in terms of their intensions and I have argued that the intension can be represented in the model as a region in Shepard's geometric PS space.

to an interval along the psychological-similarity scale, and x and y are points on this scale. On the basis of the size principle (equation 6.7), the tendency to favour generalisation towards objects that are similarly likely to belong to the same concept drops exponentially with the size of the concept that includes both x and y in a psychological similarity space (e.g., Shepard's geometric space of regions). On this basis, I have suggested a reformulation of ULG that gives candidate concepts a stronger role in shaping the generalisation gradient.

T&G (2001, p. 637) also give an alternative formulation of the generalisation function. This function takes the following form.

$$pr(h_{y\in C}|h_{x\in C}) = 1/\left[1 + \frac{\sum_{h_{x\in C, y\notin C}} pr(h,x)}{\sum_{h_{x,y\in C}} pr(h,x)}\right], \tag{9.2}$$

where

> $h_{y\in C}$ represents the hypothesis that the object y is in the extension of the concept C.
>
> $h_{x\in C, y\notin C}$ represents the hypothesis that x is in the extension of C but y is not.
>
> $h_{x,y\in C}$ represents the hypothesis that both x and y are in the extension of C.
>
> $pr(h,x)$ represents the joint probability of the relevant hypothesis (e.g., either the hypothesis that x is in the extension of C but y is not, or the hypothesis that both x and y are in the extension of C) and the evidence, x.

In equation 9.2, the weighted sums represent the totality of the probabilities associated with specific subsets of hypotheses. The connection to Tversky is that each subset of hypotheses in equation 9.2 corresponds to a subset of features in Tversky's ratio model (equation 8.1). For example, $h_{x\in C, y\notin C}$ corresponds to the features that are distinct to x but not y and $h_{x,y\in C}$ corresponds to the features that are common to x and y.[3] In this version of the Bayesian model, each subset of hypotheses is weighted according to its joint probability with the evidence, $pr(h,x)$, where x represents the evidence and h represents a subset of hypotheses. The joint probability is the product of the hypothesis' likelihood and prior. For instance, the weighted sum of all those hypotheses of the type $h_{x\in C, y\notin C}$ is a sum of each hypothesis that is compatible with x and is incompatible with y, weighted by the product of the associated likelihood and prior. The formal description of the weighting process works like hypothesis averaging (Appendix A).

[3]Note that $h_{x\in C, y\notin C}$ need not be a single element. If x is a fly agraric mushroom and y is a tree, then $h_{x\in C, y\notin C}$ may represent a set of hypotheses that pick out distinct concepts. One of these hypotheses might assign x to the concept FLY AGRARIC MUSHROOM while another might assign x to the concept EDIBLE, and neither of these concepts has y as a member.

Equation 9.2 resembles the formal structure of Tversky's ratio model. If subsets of hypotheses can be taken to represent sets of features and their corresponding probabilities represent weights, then generalisation becomes a function of the ratio of the sum of the weighted distinct features to the weighted common features, which looks like the inverse of similarity in Tversky's ratio model. On the basis of this formal resemblance, T&G (2001, p. 637) argue that their Bayesian model can accommodate the directionality in Tversky's model. I show below that this argument is underdeveloped. I refine it towards the argument that the Bayesian theory of generalisation can predict observations of the effect of directionality in generalisation, but that 'directionality' in the Bayesian model carries a different meaning from 'directionality' in Tversky's model.

A first step for this argument is to show how a change in the direction of the inference produces a change in the joint probabilities with the evidence. This would correspond to a change in the weights associated with the distinct subsets of hypotheses. To see this dependency, consider what happens when equation 9.2 is reversed, such as in the following equation.

$$pr(h_{x\in C}|h_{y\in C}) = 1/\left[1 + \frac{\sum_{h_{y\in C, x\notin C}} pr(h,y)}{\sum_{h_{x,y\in C}} pr(h,y)}\right], \tag{9.3}$$

where $h_{y\in C,x\notin C}$ is the hypothesis that y is in C but x is not. In this direction, the result of the probability function depends only on the probabilities associated with hypotheses distinct to y, whereas in the earlier direction, the result had only depended on the probabilities associated with hypotheses distinct to x.

The second step in the argument is to make explicit that directionality in the Bayesian model is a property associated with the definition of conditional probability (see Glossary). Correspondingly, the model derives directionality with regards to the two generalisation functions (equations 9.2 and 9.3) from differences between the formal properties (i.e., probabilities) associated with subsets of distinct hypotheses. On this basis, the Bayesian theory of generalisation can predict that generalisation from x to y will sometimes be differently probable than generalisation from y to x. This sense of directionality will be the case whenever the unconditional probabilities, $pr(h_{x\in C,y\notin C})$ and $pr(h_{y\in C,x\notin C})$, are different with regards to the opposite directions of the comparison. For simplification, let us abbreviate these hypotheses: $h_{x\in C,y\notin C}$ states that x is in C (in one direction) and $h_{y\in C,x\notin C}$ states that y is in C (in the other direction). Thus, directionality in generalisation derives from differences between the formal properties associated with subsets of distinct hypotheses, i.e., differences between the joint probabilities of the distinct hypotheses with the evidence. Symmetries should occur whenever the probabilities attached to these hypotheses are exactly the same, independent of the direction. This way of using the definition of conditional probability provides in some sense a 'law of directionality': given that x and y are typically different objects, it is likely that they are differently plausible with respect to a concept.

The Tel Aviv–New York example illustrates the difference between directionality in T&G's Bayesian model and Tversky's model. On a Bayesian approach, the problem faced by the subject is different from a simple similarity-judgement task. On this approach, the problem is to generalise some behaviour (e.g., travelling) from Tel Aviv to New York (in one direction) or from New York to Tel Aviv (in the other direction). To decide whether to generalise, the agent has to judge the conditional probabilities associated with these cities and some candidate concept, for example, TRAVEL DESTINATION. In one direction, the agent has to judge how plausible it is that Tel Aviv is an instance of the concept TRAVEL DESTINATION, given that New York is an instance of the concept TRAVEL DESTINATION. In the other direction, the agent has to judge how plausible it is that New York is an instance of the concept TRAVEL DESTINATION given that Tel Aviv is an instance of the concept TRAVEL DESTINATION. These conditional probabilities can be defined as follows.

$$pr(\text{Tel Aviv} \in \text{DESTINATION} \mid \text{NYC} \in \text{DESTINATION}) = \frac{pr(\text{Tel Aviv} \in \text{DESTINATION} \cap \text{NYC} \in \text{DESTINATION})}{pr(\text{NYC} \in \text{DESTINATION})} \tag{9.4}$$

$$pr(\text{NYC} \in \text{DESTINATION} \mid \text{Tel Aviv} \in \text{DESTINATION}) = \frac{pr(\text{NYC} \in \text{DESTINATION} \cap \text{Tel Aviv} \in \text{DESTINATION})}{pr(\text{Tel Aviv} \in \text{DESTINATION})} \tag{9.5}$$

A possible difference between the generalisation probabilities in equations 9.4 and 9.5 depends only on a possible difference between the probability that New York is a travel destination (in the first direction) and the probability that Tel Aviv is a travel destination (in the second direction) (see definition of conditional probability in Glossary). Intuitively, these probabilities are likely to be different because New York seems to be (on average) a much more probable travel destination than Tel Aviv.

This explanation of directionality is different from Tversky's (chapter 4) because it makes no reference to the particular differences between the contexts in each direction. We can manually add aspects of the context to the model, but these do not explain the change in generalisation probabilities. For example, it seems to be more probable that New York is a TRAVEL DESTINATION than Tel Aviv when the context is a trip to the US. On the other hand, it is more probable that Tel Aviv is a TRAVEL DESTINATION than New York when the context is a summer vacation. The appeal to the definition of conditional probability makes such directionality effects intuitively more likely. However, it does not explain how these probabilities (e.g., that New York is a more probable travel destination than Tel Aviv or vice versa) actually change with respect to a change in the context (e.g., a trip to the US versus a summer vacation). In light of the definition of conditional probability, it simply becomes more likely that generalisation is asymmetric *if* the probability of the hypothesis that Tel Aviv is in the concept TRAVEL DESTINATION is different

from the probability of the hypothesis that New York is in the concept TRAVEL DESTINATION. But the model does not say what aspects (e.g., of the context) determine the intuitive differences between these probabilities[4].

Taken together, I have shown in this section that T&G's Bayesian model is of relatively broad scope; it simultaneously predicts both instances of the ULG and instances of what may be called a 'law of directionality'. This law diverges from Tversky's explanation of directionality and it does not explain how probabilities change with the context. However, it seems to predict the same behavioural phenomena that Tversky initially took as an argument against the geometric model of PS. Thus, the Bayesian model predicts more phenomena than either Shepard's or Tversky's models were previously able to do on their own, whereby it satisfies the second criterion of unification that I have proposed. In the next section, I test whether the Bayesian theory of generalisation can also meet the third criterion of informational relevance.

9.4. Satisfying the third criterion

To recapitulate, Myrvold's (2003) criterion of unification is a reduction of the informational or probabilistic independence of two phenomena, p_1, and p_2, by accepting the unifying hypothesis, T, so that knowing T changes how positively probabilistically dependent p_1 & p_2 are on each other. In the current case, p_1 corresponds to Shepard's geometric theory of PS, including the ULG, and p_2 corresponds to Tversky's feature-matching theory of PS. p_1 obtains evidence from observations of the exponential gradient, which are denoted by e_1. p_2 obtains evidence from observations of directionality effects, which are denoted by e_2. T represents the Bayesian theory of generalisation, including T&G's size principle (chapter 6.4) and the law of directionality (section 9.3). This section argues that the Bayesian theory of generalisation satisfies Myrvold's criterion. The theory unifies the empirical instances of the ULG and of the law of directionality by making these observations look positively probabilistically dependent while these observations had previously appeared to be probabilistically independent under either Shepard's or Tversky's previous theories of PS.

To unify in this third sense, the Bayesian theory has to meet the condition that $pr(p_1|p_2, T) > pr(p_1|p_2)$ or that $pr(p_2|p_1, T) > pr(p_2|p_1)$ (cf. section 8.3.3). That is, the probability to observe the exponential gradient given directionality effects and the assumption that generalisation can be described with a function of

[4]Note that directionality cannot be a result of the likelihood if the size of the concept, DESTINATION, is fixed. Following the size principle (equation 6.7), the likelihood is a ratio of the size of the concept. If the size of the concept is fixed, then the likelihoods, $pr(x|x \in$ DESTINATION$)$ and $pr(y|y \in$ DESTINATION$)$, should be unaffected by the direction because the likelihoods are symmetric—the probabilities associated with these hypotheses must be the same if the concepts have the same size. This would suggest that directionality must depend on the prior probabilities associated with the sets of distinct hypotheses. However, T&G do not specify the prior probabilities in their model (chapter 7).

Bayesian inference must be greater than the probability to observe the exponential gradient given the observation of an effect of directionality. Alternatively, the Bayesian model unifies when the probability to observe directionality effects given the exponential gradient and the Bayesian model is greater than the probability to observe directionality effects given the exponential gradient. In the following paragraph, I focus on the contrast between T&G's and Tversky's theory, that is, on the condition that $pr(p_1|p_2, T) > pr(p_1|p_2)$.

Indeed, it seems to be relatively more probable to observe an exponential gradient given the observation of a directionality effect under the supposition that generalisation is a Bayesian inference (including the size principle and law of directionality) than when this supposition is replaced with the set of assumptions of Tversky's theory. (To understand this case, it can be imagined that the Bayesian theory plays the role of the Copernican hypothesis and Tversky's theory takes the role of the Ptolemaic hypothesis in chapter 8.3.3.) The Bayesian theory predicts (and is not only consistent with) both the exponential gradient and directionality effects in generalisation behaviour under the assumption that both types of behaviour are effects of whatever mechanism can be described with the laws of probability and the size principle (equation 6.7). In contrast, Tversky's theory fails to predict the exponential gradient—directionality effects and the exponential gradient cannot both be accommodated by the feature-matching model because the model predicts that PS is linear (chapter 4).

In this sense, the Bayesian theory renders observations of the exponential gradient positively probabilistically dependent on observations of the effect of directionality. That is, according to the Bayesian theory, the occurrence of an exponential gradient becomes more probable conditional on the occurrence of the effect of directionality than without regards to the Bayesian theory. Previously, observations of the exponential gradient and directionality effects had appeared to be probabilistically independent or even negatively probabilistically dependent under Tversky's earlier theory of PS. Therefore, the Bayesian theory of generalisation satisfies the third criterion of unification.

Taken together, I have argued in this section that the theory of generalisation inspired by T&G's model of concept learning can render the exponential gradient and the effect of directionality positively probabilistically dependent on each other.

9.5. Conclusion: unification of the phenomena

Taken together, I have argued in this chapter that the Bayesian theory unifies Shepard's (1987) observations of the exponential gradient of generalisation and Tversky's (1977) observations of directionality effects in similarity judgements by meeting the three criteria of unification that I have outlined in chapter 8. Firstly, the Bayesian theory describes these behaviours as possible solutions to a single

task of Bayesian inference (expression 9.1) under a single set of assumptions, whereas previously, these behaviours were associated with two conflicting sets of assumptions and PS functions. I have argued that this description of the behaviour is syntactically simple or elegant (section 9.2).

Secondly, I have argued that the Bayesian theory can predict both that generalisation behaviour will be exponential and that generalisation will be sometimes directional on the basis of T&G's size principle (section 9.3). Thereby, the Bayesian theory of generalisation combines the predictive powers associated with Shepard's previous ULG and the law of directionality, and predicts overall more phenomena than could be predicted in light of either the ULG or the feature-matching model alone. In this sense, the theory has a relatively broader scope than Shepard's and Tversky's theories.

Thirdly, I have argued that the Bayesian theory does more than introducing simplicity and broad scope: whereas the observed facets of generalisation (i.e., exponential and directional) behaviours appear to be independent of each other in light of either Shepard's or Tversky's previous definitions of PS, they become dependent in light of the analysis of generalisation as a Bayesian-inference problem (section 9.4). Thereby, the Bayesian theory of generalisation satisfies the third criterion of unification: it renders these types of behaviours informationally relevant to each other. Thus, overall, the Bayesian theory of generalisation inspired by T&G's (2001) model of concept learning earns the label of a unifying theory.

My argument in favour of unification plays out positively for a strengthening of T&G's (2001) Bayesian approach to generalisation and concept learning in light of the reverse-inference scheme introduced in chapter 1.

Argument structure of T&G's (2001) reverse inference:

(A) When psychological process, CL, is recruited by a Bayesian-inference task, an exponential gradient of generalisation, E, *and directional generalisation, E'*, are likely to be found.

(B) In Bayesian-inference task T, E *and E'* were found.

(C) Hence, psychological process, CL, was recruited by Bayesian-inference task T.

Here, the additional evidence, E', in (A) and (B) seems to lend further support for the conclusion, (C). Instead of making one observation (E) more probable, the Bayesian approach makes two observations (E and E') more probable at the same time. Thus, in combining the predictions from Shepard's and Tversky's approaches, the Bayesian theory picks up more support by the total evidence (E plus E'), and seems to be empirically better confirmed.

The proposed unification is a unification of the empirical phenomena of exponential generalisation and directionality effects from a computational level perspective. From this perspective, both of these types of behaviours can be analysed

and predicted in terms of a problem of Bayesian inference. It is not a unification of the conflicting theoretical assumptions that are associated with Shepard's (1987) and Tversky's (1977) earlier theories of PS.

This clarification motivates two points of reflection. Firstly, the choice of the Bayesian theory of generalisation over either the geometric or the feature-matching theories is a choice at the computational level of explanation. At this level, the particular psychological processes that have generated the behavioural patterns are still unknown. For example, the behaviour might be generated by a process of concept learning (as proposed by T&G, 2001), a process of PS (as proposed by Shepard, 1987, and Tversky, 1977) or a combination of these. When asking about the possible psychological mechanisms that have generated the associated observations of the exponential gradient and the effect of directionality, a decision has to be made about which of Shepard's and Tversky's conflicting sets of assumptions (chapter 5) should be taken on for further investigations at these levels. In other words, the competition between the sub-theories about questions at this level of explanation still remains.

Secondly, at the computational level of explanation, the advocated Bayesian approach is a third alternative approach to explaining PC. The proposed unification of the phenomena by the Bayesian theory of generalisation concerns a single set of theoretical assumptions (e.g., a hypothesis space, a prior probability distribution over hypotheses and a likelihood function that relates the evidence to the probability distribution) that is distinct from the sets of assumptions associated with Shepard's and Tversky's previous theories. This can be seen most clearly in the contrast between the law of directionality that is implicit in T&G's model and Tversky's set-theoretic and psychological assumptions about the asymmetry in similarity judgements. It can also be seen with regards to the differences between T&G's and Shepard's distinct assumptions about the sampling process that learners use when they infer the most probable concept (i.e., either weak or strong sampling), which I have explained in chapter 6.

In light of these reflections, the Bayesian unification of the phenomenon of PC does not replace the possible representational- and algorithmic level explanations of PS offered by Shepard's and Tversky's theories. The next chapter concludes with the implications of the proposed unification for the problem of modelling PC and open questions.

10. Conclusion

10.1. Recapitulation of the problem and the claim

This thesis has addressed the problem of modelling perceptual categorisation (PC). To recall, the problem was to explain our ability to generalise behaviour from old perceptual experiences to new perceptual experiences. For example, upon having eaten an umami portobello mushroom and in light of the new experience of a bitter fly agraric mushroom, we will be likely to seek further instances of portobello mushrooms and avoid instances of fly agraric mushrooms. I have addressed this problem from a computational perspective.

Building on this perspective, my claim in this thesis was that PC can be modelled as a unified phenomenon with a Bayesian approach that is inspired by Tenenbaum and Griffiths' (2001) model of concept learning. From the perspective of this approach, the ability to generalise behaviour from old perceptual experiences to new perceptual experiences can be analysed as a problem of Bayesian inference at the computational level: given the observation of an umami portobello mushroom, an ideal agent can evaluate the plausibility of a hypothesis about what candidate concept or category (e.g., EDIBLE) this mushroom belongs to with Bayes' Theorem. If this inference strategy reveals that the most plausible category is different with regards to the observation of the bitter fly agraric mushroom (e.g., if it turns out that the portobello mushroom belongs to the category of edible mushrooms but the fly agraric does not), then the agent should be likely to avoid eating the fly agraric mushroom. From the Bayesian perspective that I have proposed here, the unified explanation of why the agent is likely to display such behaviour is that the behaviour can be accurately described and predicted with the principles of Bayesian inference and additional assumptions about the origin of the perceptual observations (e.g., assumptions about how the objects have been sampled).

10.2. Summary of the overall argument in support of the claim

I have supported my claim that the Bayesian approach can unify the phenomenon of PC in chapters 8 and 9. There I have argued that my approach to PC meets three philosophical criteria of unification: (i) it can elegantly describe and (ii) predict two aspects of PC behaviour (the exponential gradient and the effect of directionality), and (iii) make these aspects appear relevant to each other.

My approach builds on the background of Marr's (1982) three-levels of analysis, Anderson's (1991a) rational analysis and their combination within a reverse-engineering strategy in cognitive science. I have used these frameworks to position my unifying approach to PC at the computational level of explanation (chapter 2). At this level, the target is to explain what the agent's problem is (e.g., to infer the category that the perceptual experience belongs to), why this problem is appropriate for the agent to have (e.g., it helps the agent to thrive and survive), and what the principles are that guide a possible solution to this problem (e.g., the agent should prefer relatively smaller categories over larger ones/the size principle). I have argued that the approach is idealised at the computational level but it can be combined with Shepard's (1962; 1987) theory of psychological similarity, in which concepts are represented as geometric regions. Shepard's theory offers one way in which the concepts that constitute the contents of hypotheses in the Bayesian model could be individuated at the level of representation and algorithm (chapter 6). Thus, the proposed unification contrasts with Tenenbaum and Griffiths' earlier approach in that it provides a philosophical argument for unification and it makes the connection between Tenenbaum and Griffiths' Bayesian and Shepard's similarity-based approaches explicit. Previously, no argument for unification had been provided and the connection between these approaches was implicit.

To sum up, here is what I have done to reach this conclusion. In Part I of this thesis, I have discussed the key contributions of Shepard's (1987) and Tversky's (1977) explanations of two types of behaviours associated with PC. These behaviours are (i) the exponential gradient of generalisation, which Shepard had described and predicted with the Universal Law of Generalisation (chapter 3), and (ii) the effect of directionality in similarity judgements, which Tversky could accommodate with the feature-matching model (chapter 4). In chapter 5, I have contrasted these two competing approaches with respect to their conflicting assumptions about the kinds of psychological similarity representations and processes that explain these observed behaviours (i.e., assumptions about psychological similarity spaces, explanatory targets, structures of mental representations and interpretations of the empirical data). This contrast, but also their common problem of explaining behaviour that appears to be a kind of PC, has motivated my project of unifying the central observations (the exponential gradient and the effect of directionality) in Part II of this thesis.

In Part II of this thesis, I have supported my claim that PC can be studied as a unified phenomenon. I have advocated a Bayesian theory of generalisation that is inspired by Tenenbaum and Griffiths' (2001) Bayesian model (chapter 6). In chapter 7, I have argued that this theory explains observations associated with PC behaviour at the computational level of analysis. I have subsequently argued for criteria (i)-(iii) of unification (chapter 8) and shown that the Bayesian theory of generalisation inspired by Tenenbaum and Griffiths' model can meet these criteria (chapter 9). Therefore, PC can be studied as a unified phenomenon at the computational level of analysis.

Three lessons have been learned from this thesis. Firstly, although Shepard's and

Tversky's approaches to psychological similarity compete with regards to their assumptions about the underlying psychological spaces and representations in subjects' minds, these approaches address a common question: why does the apparent 'form' of psychological similarity have the 'shape' that it does (e.g., negatively exponential and directional)? Shepard's and Tversky's answers to this question point to different candidate psychological processes that may be associated with the actual computation of a psychological similarity function (e.g., a multidimensional scaling algorithm, the contrast model or the ratio model). Secondly, my proposed unification shows that we can stay neutral on the Shepard-Tversky debate while offering a comprehensive 'picture' that includes either of these candidate psychological processes as possible explanatory alternatives. This approach is useful because it combines the predictive powers of Shepard's and Tversky's models, it is simple, and it lets the different forms of PC behaviour appear to be compatible with each other, while they had appeared to be incompatible before. Thirdly, my approach to the problem of modelling PC is a unification of the *phenomena* of the exponential gradient and the effect of directionality. Because the Bayesian theory of generalisation does not unify the conflicting sets of theoretical assumptions of Shepard's and Tversky's theories, it is a third theoretical alternative for explaining PC.

These lessons indicate the boundaries of my proposal. In particular, my reply to the Shepard-Tversky debate does not fuse their distinct answers regarding the structure of mental representations and possible algorithms into a single answer about what the psychological space of representations and algorithms looks like. Therefore, I have not 'resolved' the Shepard-Tversky debate when asking about the representations and algorithms that instantiate the task of PC. When asking questions at the level of representation and algorithm, the competition between Shepard's and Tversky's theoretical assumptions remains.

What are the implications of my approach? I position my approach at the computational level of explanation. Therefore, it is possible that my approach competes with only one of the two other competing approaches at the representational and algorithmic level. In fact, when I see my proposal as a third way of approaching PC, then this third way can still be connected to one of the alternative ways of approaching the problem—the Bayesian approach need not exclude both of these theories. A possible connection between these approaches depends on which of the two competing theories at the representational and algorithmic levels are most coherent with the proposed unifying theory of the phenomena.

In this thesis, I have already addressed aspects of coherence and incoherence between Tenenbaum and Griffiths' unifying theory of generalisation and either Shepard's or Tversky's competing theories of psychological similarity. In parts of chapters 6 and 9, I claim that the Bayesian theory of generalisation inspired by Tenenbaum and Griffiths is intuitively more coherent with Shepard's approach to psychological similarity. I have offered two reasons for this claim. Firstly, both theories can be identified with a reverse-engineering approach to PC (chapters 3 and 6). Secondly, the mushrooms example of generalisation, which I have used throughout this thesis, shows that Tenenbaum and Griffiths' analysis implies

aspects of Shepard's analysis of the generalisation problem (chapter 6). Both approach the problem as one of inferring the most probable concept. The difference is that they make different assumptions about the principles that guide this inference, and Shepard makes additional assumptions about how concepts are represented in psychological space. In contrast, the Bayesian theory has a weaker theoretical connection to Tversky's theory of directionality because these theories refer to a different sense of 'directionality' when they are used to explain why generalisation and similarity judgement tasks are sometimes directional (chapter 9). The Bayesian theory considers directionality to be governed by a law of probability but Tversky's theory considers directionality to underlie the constraints of set-theory.

10.3. Questions for future work

Taken together, these lessons and implications provoke two questions for future research.

1. Is the proposed Bayesian unification better than either Shepard's or Tversky's theories at explaining PC?
2. Which of Shepard's and Tversky's competing theories should be chosen to investigate possible PC mechanisms at the level of representation and algorithm?

An answer to the first question seems to depend on the level at which PC should be explained. At the computational level, there are reasons to believe that the unifying Bayesian theory is better when contrasted with both Shepard's and Tversky's theories. This is because the Bayesian unification seems to be better confirmed by the total evidence than these competitive theories at the representational and algorithmic level. I have illustrated this with the reverse-inference scheme in chapter 9, where it seems that the Bayesian theory picks up relatively more evidential support by the conjunction of the observations of the exponential gradient and the effect of directionality. The Bayesian theory does this in combining the relevant predictions from Shepard's and Tversky's models, and by rendering their observations positively probabilistically dependent. However, when questions about the internal structure of mental representations in the Shepard-Tversky debate become important, the Bayesian unification (at least in isolation) does not seem to address these questions better than either of Tversky's or Shepard's theories. The approach is agnostic about how PC can be explained at the level of representation and algorithm. The Bayesian approach does not specify how hypotheses can be individuated with regards to the intensions or extension of concepts (chapters 6 and 7).

A choice in favour of one over the other approaches across these levels seems to be difficult under the supposition of Marr's framework. This framework is useful to

communicate my approach but it is also somewhat idealised (Danks, 2008).[1] For example, aspects of Shepard's theory can be positioned at the representational and algorithmic level and also at the computational level. At the computational level, Shepard's and Tenenbaum and Griffiths' theories compete with regards to their conflicting assumptions about the sampling process. These assumptions provide answers to the question of why the generalisation problem is appropriate given the relevant adaptive or communicative constraints on the agent (chapters 6 and 7). However, Shepard's and Tenenbaum and Griffiths' theories are simultaneously compatible with regards to their assumptions about the structure of psychological space (chapter 6). Likewise, Tenenbaum and Griffiths' theory competes with Tversky's assumptions about the principles of directionality (chapter 9), however, these models are principally compatible with respect to the structure of concepts: there is nothing in the Bayesian approach that prevents hypotheses in $\mathcal{H}$ from being individuated in terms of discrete sets of features. In light of these preliminary considerations, the very clear contrast between these theories in terms of the levels-distinction is idealised. It is one future project of mine to identify the relations of coherence between Shepard's and Tenenbaum and Griffiths' and between Tversky's and Tenenbaum and Griffiths' theories beyond the levels-distinction.

An answer to the second question will depend on the connections between Shepard's and Tversky's competing approaches and Tenenbaum and Griffiths' unifying theory and the relations between each of these three theories and the empirical evidence. In the context of this thesis, there seems to be a greater coherence between Tenenbaum and Griffiths' Bayesian theory of generalisation and Shepard's geometric theory of psychological similarity. This is apparent in light of the mushrooms example from chapter 6 and Machery's (2013) reverse-inference scheme that I have used to explain the argument patters in these approaches. Intuitively, if the Bayesian theory is overall better confirmed than either of Shepard's or Tversky's theories (because it unifies their predictions), then the greater coherence with Shepard's theory could provide a reason for thinking that a Bayesian approach to PC will indirectly add further theoretical confirmation for Shepard's model over Tversky's model.

A future project of mine is to test this intuition, and to show that the Bayesian theory of generalisation in fact constrains the relative degrees of confirmation of Shepard's and Tversky's theories. The motivation for this future project is to investigate whether the intuitively greater coherence between Shepard's theory and Tenenbaum and Griffiths' unifying theory and the relatively small coherence between Tversky's competing theory and Tenenbaum and Griffiths' unifying theory is enough reason to choose Shepard's over Tversky's model of psychological similarity for further investigations on the possible mechanisms underlying PC.

[1] Danks (2008) has already started to develop an alternative idea of the levels that is orthogonal to its normal use. Danks argues that in principle, each of the levels can be applied at each level in this cascade, so that the hierarchical interpretation of the levels distinction is not the only plausible interpretation.

One way in which this could be done is on the basis of the Bayesian network analysis by Colombo and Hartmann (2017, pp. 471–480). Their analysis illustrates that, under certain conditions, the additional evidential support for a unifying theory lends indirectly additional support for either of two unified theories. Two of these conditions are that (1) the sub-theories must compete with each other and (2) the unifying theory must have a positive connection to one of the unified theories but, simultaneously, a negative connection to the other[2]. Colombo and Hartmann (2017) argue that, if these conditions are met, then the unifying theory may offer a constraint on the relative degree of confirmation of the sub-theories. Thereby, it becomes possible to make a choice between the unified theories when one but not the other is relatively better supported by the unifying theory.

I end this chapter with a chart that summarises my conclusion and maps the territory for future directions.

[2]Colombo and Hartmann (2017) specify these connections formally as correlations between nodes that represent these theories in a Bayesian network. A positive correlation indicates a strong coherence between the unifying theory and a sub-theory, while a negative correlation indicates a weak coherence between these theories.

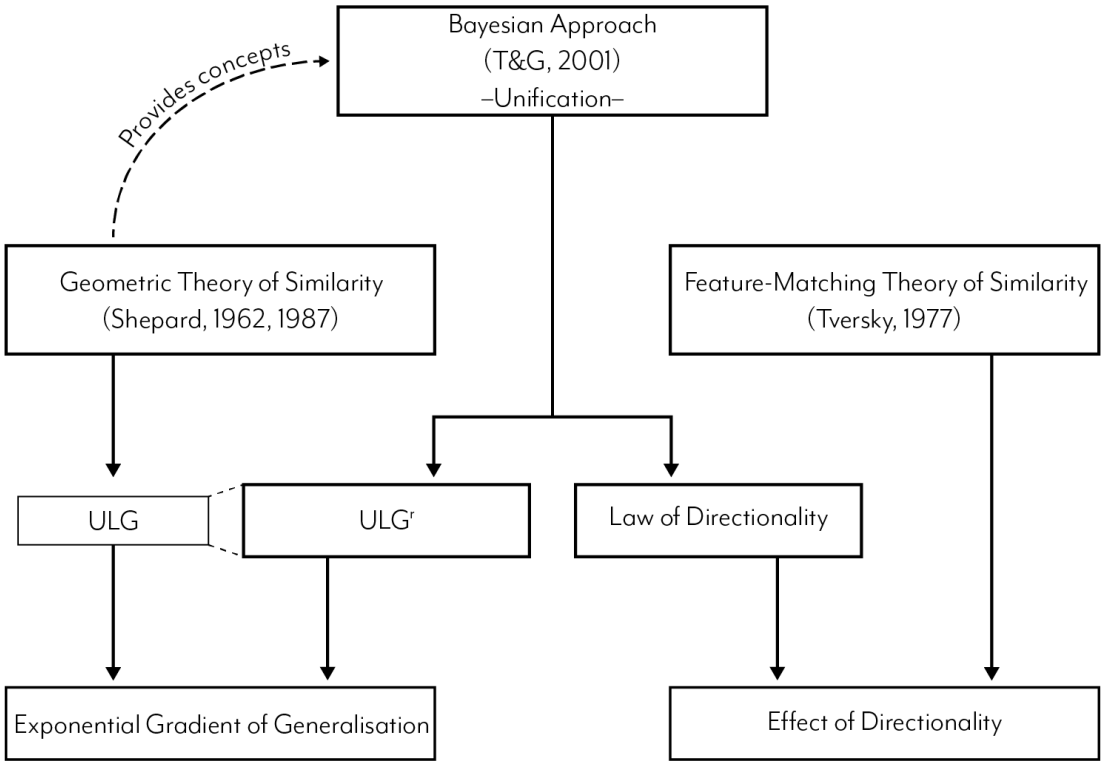


Figure 10.1.: Illustration of the position of the proposed unification with regards to Shepard's (1987), Tversky's (1977) and Tenenbaum and Griffiths' (2001) approaches. The chart illustrates the closer intuitive connection between Shepard's geometric theory and Tenenbaum and Griffiths' Bayesian approach on the upper left, where Shepard's geometric regions provide the concepts for the Bayesian model. On the right hand side, there is no connection to Tversky's theory (although a connection could be added in the future). On the bottom left, it is indicated that Shepard's theory predicts the exponential gradient on the basis of the Universal Law of Generalisation, which is generalised towards the generalisation of the Law of Generalisation with the size principle (indicated by the dotted lines). In the bottom middle, it is indicated that Tenenbaum and Griffiths' model predicts the effect of directionality, but does so with a 'Law of Directionality', which does not resemble the assumptions in Tversky's set-theoretic explanation of these effects. Thus, together, we see that the Bayesian approach is powerful in unifying the phenomena associated with PC behaviour and is more closely connected to Shepard's theory of psychological similarity. Many thanks to Nicholas Rebol, who implemented the design of this chart.

Appendix

A. Hypothesis Averaging

Hypothesis averaging is a part of the third ingredient of Tenenbaum and Griffiths' (2001) Bayesian model of concept learning in chapter 6.3. Hypothesis averaging is a (despite not the only) method for selecting hypotheses in Bayesian inference. How does it work? Roughly, a subset of the most probable hypotheses is selected and each hypothesis is weighted by its posterior probability. The resulting posterior distribution combines the predictions of several plausible hypotheses. In T&G's model (equation 6.1), the degree to which an agent generalises behaviour from x to y is calculated as the average of "the predictions that each individual hypothesis makes about y's membership in C, weighted by the posterior probability[, $pr(h|e)$,] of that hypothesis." (Tenenbaum & Griffiths, 2001, 631) In other words, equation 6.1 computes the average of the posterior predictions, $pr(y \in C|x \in C)$, considering each of the available hypotheses, $h_1, ..., h_n \in \mathcal{H}$, weighted by their posterior probability, $pr(h|e)$.

Implicit in this formulation is that the hypotheses represent ways in which a psychological space of object representations such as x or y can be carved up into sets that include either only x or only y, or both x and y—each hypothesis pairs x or y, with a candidate concept. The selection of the relevant hypotheses is an issue that depends on the task. In the current case, the task is to predict whether y belongs to the same concept, C, as x. The relevant hypotheses in $\mathcal{H}$ are those that are compatible with y. These are hypotheses of the form $h : y \in C$ (i.e., the hypothesis states that the instance y is in the intension/extension of the concept C).[1] (If we were interested in the posterior predictive distribution associated with the converse claim, we would instead specify the posterior distribution associated with the subset of those hypotheses that pick out regions that do not contain y.) There are many hypotheses of the form $h : y \in C$ because there are many possible regions that contain y. Take the hormone-levels case. On a one-dimensional scale, C_1 could be the interval covering all levels between 40 and 80, and C_2 could be the interval covering only levels from 50 to 80, and so forth. In general, there are many hypotheses whose predictions seem plausible, $h_1 : y \in C_1, h_2 : y \in C_2, h_3 : y \in C_3$, ..., etc. Hypothesis averaging is a nice method because no strict selection has to be made. The posterior predictive distribution takes the predictions of all of these hypotheses into account, weighted by their relative importance (posterior probability).

On T&G's (2001) Bayesian account, the posterior predictive distribution reflects subjects' generalisation behaviour. T&G's model makes the following empirical predictions. If the sum of the probabilities associated with the subset $h : y \in C$

[1] 'Compatible' here means that the hypothesis indicates a concept that contains both x and y.

is greater than the sum of the probabilities associated with the subset $h : y \notin C$, then subjects should generalise their behaviour from x to y. This tendency to generalise should become proportionally stronger with a relative increase in the difference between the sums of these probabilities. That is, if the sum of the probabilities associated with $h : y \in C$ is much greater than those associated with $h : y \notin C$, subjects should strongly generalise from x to y. If there is hardly any difference between these distributions, subjects should display indifferent generalisation behaviour. For instance, they should be equally likely to generalise from x to y as to not generalise from x to y.

Glossary

The following are working definitions of the central concepts that I use throughout the monograph.

Axioms of probability (also called 'Kolmogorov's laws of probability'). These laws or axioms state that for a closed set of subsets in a space of propositions or events, $A, B \in \Omega$, (1) the probability of an event or proposition is a value in the interval between 0 and 1, that is $0 \leq Pr(A) \leq 1$, (2) the probability of any event to happen or proposition to be true is 1, $Pr(\Omega) = 1$, and the probability of either event or proposition, if they mutually exclude each other, is the sum of the individual probabilities associated with each of them, $Pr(A \cup B) = Pr(A) + Pr(B)$, if $A \cap B = \emptyset$.

Bayes' Theorem. Bayes' Theorem is a principle of reasoning about a hypothesis, H, in light of some evidence, E. The Theorem has been proposed by famous Reverend Thomas Bayes (ca. 1701-1761) and is formalised as follows.

Theorem 1 (Bayes' Theorem)**.** *The probability of a hypothesis given a piece of evidence is equal to the ratio of the product of the hypothesis' likelihood and the prior probabilities to the probability of the evidence.*

$$Pr(H|E) = \frac{Pr(E|H) \times Pr(H)}{Pr(E)}$$

Where '$Pr(H|E)$' is the posterior probability of the hypothesis given the evidence, '$Pr(E|H)$' is the likelihood and '$Pr(H)$' is the prior probability of the hypothesis regardless of the evidence.

Bayes' Theorem can be derived with the product rule, which is the same in both directions, so that $Pr(H|E) \times Pr(E) = Pr(E|H) \times Pr(H)$. Dividing by $Pr(E)$ on both sides obtains Bayes' Theorem: $Pr(H|E) = Pr(E|H) \times Pr(H)/Pr(E)$.

Bayes' Theorem has found many applications in reasoning and inference problems. For example, a doctor might use Bayes' Rule to estimate the probability that a patient has one of three possible disease given that she is coughing. The three possible disease are three hypotheses, H_1, H_2 and H_3 and the coughing is the evidence, E. H_1 says that the patient has a cold. H_2 says that the patient has heartburn and H_3 says that the patient has lung cancer. Following Bayes'

Theorem, each hypothesis is evaluated based on its likelihood and prior probabilities. H_1 and H_2 obtain the highest likelihoods because it happens very often that someone who actually has a cold coughs and also that someone who actually has lung cancer coughs. In contrast, H_1 obtains only a small likelihood: it is unlikely to observe someone coughing given that she only has a heart burn. The prior probability for H_2 is much smaller than for H_1. It is much less probable for the patient to have cancer than to have a cold, regardless of whether she is coughing or not. Taken together and following Bayes' Theorem, H_1 will win above the other candidate hypotheses and the doctor can infer that given the evidence and background knowledge about all possible disease, it is most plausible to believe that the patient has a cold.

Bayes' Rule. This is a learning rule that is commonly used to combine the information about the data that is encoded in the likelihood and the prior probability to update a new posterior probability. In cognitive science and epistemology, Bayes' Rule is typically used to dictate how a rational agent should revise her existing beliefs in light of novel evidence.

Definition A.0.1 (Bayes' Rule)**.** *If $Pr_{old}(H)$ is an agent's old degree of belief in a proposition H and E is a new piece of information with $Pr(E) > 0$, then the agent's degree of belief in H should change to Pr_{new}, which is the agent's old degree of belief in H conditional on E. Formally: $Pr_{new}(H) = Pr_{old}(H|E)$.*

Definition A.0.1 is also referred to as *strict conditionalisation* (e.g. Brössel, 2015; Huber, 2016). The underlying assumption is that probabilities measure a rational agent's degrees of belief in the truth of a hypothesis (typically interpreted to be a proposition). A learner is rational if and only if the structure of its belief system satisfies a condition of probabilistic coherence over time.

Category. This is a group of things that share one or more properties. For example, the category *dog* is the group of things that typically share properties such as *being four-legged*, *being furry*, *being smelly*, etc. This thesis focuses on perceptual categories, which are based on perceptual properties, as opposed to unperceptual categories, like *democracy* or *good*. In the contemporary literature, it is much disputed whether categories are ontologically real or whether they are mind-constructed groupings.

Under the cognitive/mental conception of categories, the explanatory target of a theory of PC is to explain the cognitive processes that are at work in carving up perceptual representations of the world.

Cognitive-computational model. Generally, a model is a representation that provides a structure for identifying a target phenomenon. The relevant structure is a set of states and a set of transitions between them. In the case of a

cognitive-computational model, the represented structure is a cognitive process. A computational model of some aspect of cognition provides a description of the transformation of a set of input states into a set of output states. For instance, a computational model of perception provides a description of a set of perceptual states as inputs, a description of a set of actions as outputs and a description of the transformations of perceptual inputs into action outputs. Such a model might include the description of the rules that govern the transformations in the computational process. For instance, a model of action might use the rule: given the option of tea or coffee and knowing that the work meeting lasts three hours, drink the coffee because it increases the chances of working better. More generally, the rule says that given a set of perceptual input information pieces and in light of background information about the situation, choose the action with the highest utility.[2]

Cognitive-computational models are useful instruments for understanding cognitive functions, which cannot be directly observed by scientific study (e.g., by neuroscientific or behavioural methodology). Sun (2008) lists a variety of positive aspects that come with computational cognitive modelling more generally.

1. Computational cognitive modelling is a tool for understanding the mind, where this level of understanding is impossible to achieve given an analysis of behavioural data alone because the latter is just counting correlations.

2. It is a tool for discovering inconsistencies in the theoretical assumptions that experiments are usually performed on – usually common-sense.

3. It is a tool for bringing out fine details in a process and thereby helps to gain more conceptual clarity and explanatory precision.

4. It is a tool for thought experiments and hypothesis generation.

5. Though this has according to Sun not yet been successfully implemented, computational cognitive models might in the future help to unify superficial explanations across multiple domains (in analogy to Einstein's theory that unified electromagnetic and gravitational forces). It may also contribute against the increasing overspecialisation that may otherwise prevent communication across disciplines

[2] A computational model is different from a mathematical model in that it is capable of representing a process at a more detailed level than a mathematical description of the function that could generate such a process would do (Weisberg 2012, ch. 3; Sun 2008, p. 4). The purpose of a computational model is to specify the process by which the inputs are transformed into output states. A mathematical model is broader in that it specifies those ways of formally representing cognitive phenomena which do not focus on a process or a procedure but represent static relations between entities. An example is the mathematical Lotka-Volterra model (Volterra, 1926), which represents the proportion of predator and prey in a given state. Thus, computational-cognitive modelling is a way of representing cognitive *processes* more than individual cognitive states and other more general aspects of cognitive phenomena.

Although different from a computational theory, a computational model is often a helpful place to start out with an explanation. In this sense, a theory of a cognitive process stands to a cognitive-computational model like an explanans to its explanandum.

For example, perception is a cognitive process that is difficult to directly observe. By creating a computational model of perception, it is possible to give an example of how perception could work, given a set of parameters and theoretical background assumptions. For instance, if one assumes that perception is a Bayesian inference (cf. Knill & Richards, 1996), one can construct a model of perception that uses a Bayesian inference algorithm to infer hidden causes from inputs in the world (Clark, 2013). This information can be encoded in weighted nodes of a neural network that should update these parameters upon novel incoming information in line with Bayes' Rule. Predictions of the model can then be compared to interpret experimental results. Thus, computational-cognitive modelling can offer possible explanations of a phenomenon.

Computational models should accompany behavioural experiments. It is difficult to find out which cognitive process was relevant in producing the observed behaviour from observations of generalisation behaviour alone. Computational cognitive modelling can help to limit the space of scientific hypotheses about what cognitive process could be relevant in generating the generalisation behaviour.

This thesis investigates a couple of computational-cognitive models of PC that may offer adequate constraints on a theory of PC that explains how the psychological mechanism of PC works. Chapter 1 refers to one popular view on the levels of explanation along which computational cognitive modelling can be characterised – Marr's (1982) three levels of explanation of an information-processing system.[3]

Cognitive-computational theory. This is an explanation of a cognitive phenomenon in terms of computation. A cognitive-computational theory is different from a cognitive-computational model in that the former is an explanation while the latter is a representation of a cognitive phenomenon in terms of computation. Cognitive-computational modelling is useful for cognitive-computational theorising, e.g., when direct experimenting on the brain is not practicable or when the explanatory target is highly complex (cf. Stinson, 2018, p. 121). One instance of a cognitive-computational theory is the computational theory of mind (CTM). Roughly, CTM puts forward two central claims: (1) psychological phenomena may be explained mechanistically (A. Isaac, 2018) and (2) thought is a kind of computation (Casey & Moran, 1989). The goal of CTM is to explain how the mind works based on the assumption that the mind works like a computing system.

[3] Danks (2008) and Newell (1994) offer alternative analyses of levels of explanation in cognitive science.

CTM is inspired by Turing's (1936) idealised model of the mind. The classical version of the theory assumes that thinking can be realised in a machine in which symbolic computations are governed by routine mechanical instructions. As discrete symbol-manipulation may not be representative of human thinking, novel interpretations of CTM consider other computational operations, such as analogue computations. These seem more plausible with respect to sensory processing (Shu, Hasenstaub, Duque, Yu, & McCormick, 2006)[4]. Because analogue computations are different from the idealised Turing model, it is not yet clear how these computational approaches figure within a mechanical framework.

Concept. There are roughly three major camps in the philosophy of mind as to what a concept is. On the Fregean view, concepts are abstract entities, *senses*, which determine the semantic content of statements and do not differ in their truth-values or objects. For example, the difference between "the morning star is the morning star" and "the morning star is the evening star" is not in their truth-value; both refer to the same object (i.e., Venus) but they each have a different sense.

The Fregean view contrasts with a common view in philosophy of cognitive science, where a concept is a mental representation that figures into propositional attitudes (mental representations about states of affairs that have a causal-functional role for producing behaviour directed towards these states of affairs) like atomic symbols do in sentences (Fodor, 1975, 1987; Margolis & Laurence, 1999, ch. 1). On this view, which has been dubbed 'Representational Theory of Mind' (henceforth 'RTM'), concepts fulfil a meaning-constitutive role for propositional attitudes (e.g., beliefs, desires, ...). For example, the belief that 'Paris is the capital of France' is composed of the concepts PARIS, CAPITAL and FRANCE. The way in which these concepts are arranged 'composes' the meaning of the belief that Paris is the capital of France, or, in other words, the content that this believe represents at the personal-level of cognition. According to RTM, concepts are individuated (i.e., physically realised) in the brains of those who think/represent such a belief. Many different versions of RTM have been proposed and many of them are not committed to the idea that concepts are symbolic representations. It is commonly advocated that concepts can be represented in a variety of different formats, e.g., in terms of imagistic (cf. Kosslyn, 1980) or map-like structures (cf. Blumson, 2012; Camp, 2007) and in subpersonal-level theories of cognition (cf. Von Eckardt, 2012). The Fregean and the mental-representations views are principally not incompatible. This has been outlined by Margolis and Laurence (1999, p. 8), who suggest that the individuation conditions of concepts as mental representations are also partly determined by a Fregean sense.

A third approach is to define a concept as an ability or a normal function (Millikan, 1989). For example, the concept APPLE is whatever capacity enables an organism to discriminate apples from non-apples and to infer the relevant consequences from experiences with apples, e.g., to infer from eating an apple that

[4]For a critical perspective on the analogue-discrete distinction, see Maley (2011)

apples are edible. Following the ability-view, mental states carry meaning in terms of the purpose that they fulfil for the representing system (e.g., attainment of nutrition and survival). Millikan (1989) gives the example of a frog's ability to identify (viz., have a concept of) a fly: the capacity of the frog's visual system to produce a representation of the fly is grounded in the need of the frog's co-evolved stomach to digest nutrients. The rationale in this explanation is consumer oriented: aspects of the consumer system explain aspects of the visual system. Mental content (e.g., 'fly') is determined by natural selection of a producer system (e.g., the frog's visual system) for a consumer system (e.g., the frog's stomach). The producer system (i.e., the visual system) is the representational vehicle and the consumer system (i.e., the stomach) is the target as it is the device that uses the representation. From Millikan's perspective, what it is to have a concept is to have an adaptive function (e.g., perceiving x) for a device (e.g., a stomach) to carry out its capacity (e.g., digesting x).

Throughout this thesis, I follow the typical understanding in cognitive science with a twist towards the ability-view. I use the word 'concept' to refer to an agent's mental representation of possible perceptual inputs, where this representation enables the agent to categorise those inputs as belonging to the same or a different category. For example, the concept DOG is a mental representation of whatever could be a dog-like perceptual experience. Such a mental representation represents not only information about experiences that are associated with actual perceptions of dogs, but also ways in which something that would be categorised as a dog, were the agent to perceive it, could be experienced. For example, even though I may have never seen a Boxer-Pitbull Mix, my DOG concept must include an instance of this breed as a possibility so that upon seeing a Boxer-Pitbull Mix, I will be able to use that concept to categorise it as a perceptual experience of a dog. Thus, a concept represents information about past but also possible future perceptual experiences (e.g. Boxer experiences and Boxer-Pitbull Mix experiences), which require appropriate actions (e.g., not stroking the dog) in novel situations.

Given this working definition, a concept is a mental representation that enables one to categorise perceptual experiences of objects or situations in a way that leads to efficient generalisation of behaviour across the individual experiences associated with the particular objects or situation. In this sense, concepts are equivalent to cognitive categories (in Rosch's sense), however, not to categorisation.

This thesis focuses on *perceptual* concepts such as DOG, GREEN or MILK, and largely disregards *abstract* concepts like DEMOCRACY, ELECTRON or INTERNET. I endorse the possibility that the cognitive mechanism of perceptual-concept individuation may be independent of the cognitive mechanism responsible for abstract-concept individuation. Concepts may be perceptual or linguistic entities with a variety of different representational formats. Their semantic content may be encoded in continuous perceptual representations or in discrete linguistic symbols (cf. Gärdenfors, 2000, 2014).

Concept learning. This is a cognitive process by which an agent acquires a mental representation from her experience in light of some background knowledge and with an inductive-inference strategy. For example, upon a few perceptual experiences of mushrooms that the agent categorises as 'yummy', she forms a mental representation that pairs the word 'yummy' with (a) information about her past perceptual experiences associated with the word 'yummy' and (b) information about those things that the agent believes should be labelled 'yummy' in the future. The process is indeterminate as many interpretations of the meaning of 'yum' are compatible with the paired association of the mushroom experiences (cf. Quine, 1960). The possibility of concept learning has been challenged by Fodor (Fodor, 1975, 1998; Stöckle-Schobel, 2012).

Conditional probability. The probability of an event to occur or a proposition to be true can be conditionalised upon the assumption of another event or proposition. For example, my degree of belief that it is going to rain tomorrow is very high. But suppose the weather forecast has announced a sunny day tomorrow. Given this information, my degree of belief that it is going to rain tomorrow would go down. This is a conditional degree of belief.

Definition A.0.2 (Conditional probability)**.** The probability that one proposition is true given another proposition is true. The probability of one proposition, A, conditional on another proposition, B, is a ratio of their joint probabilities and the unconditional probability of B. Formally:

$$pr(A|B) = \frac{pr(A \cap B)}{pr(B)}.$$

The relationship is directional, so that, conversely,

$$pr(B|A) = \frac{pr(A \cap B)}{pr(A)}.$$

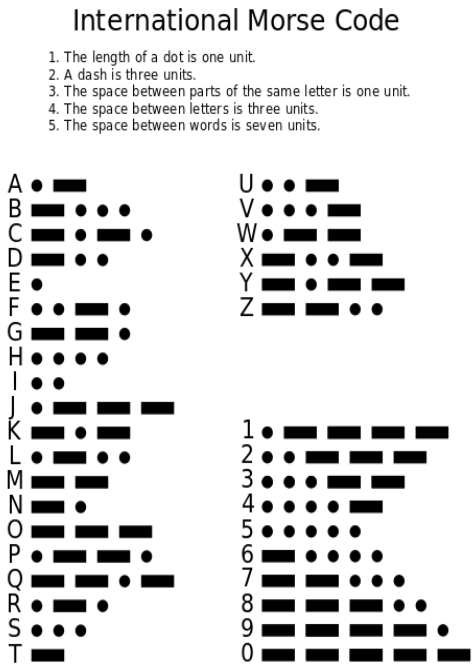
Confusion probability. In psychophysical experiments, this is the probability of a subject to wrongly identify two stimuli as the same. Typically, this is measured as the percentage of errors associated with "same" or "different" judgements for a pair of stimuli. For example, if out of 100 trials, a pair of stimuli, $\{a, b\}$ is judged to be the same in 40 trials, then the confusion probability associated with that pair and at position is 40%, or '04'.

Generic function. A function that is characteristic of a mapping is a generic function. This function does not provide precise information about how the mapping is achieved. Descriptions of functions at the computational level of explanation in cognitive theories (see 'Cognitive-computational model') are often

interpreted in terms of generic functions. At the computational level, the function specifies the abstract form of the mapping between a set of inputs and a set of outputs, but does not characterise the rules following which the mapping between inputs and outputs can be achieved. The rules are specified at the level of representation and algorithm (Colombo & Seriès, 2012, 698).

Hypothesis space. In Bayesian inference models of cognition, the agent represents uncertainties associated with events in the world by probability distributions (Wiese, 2017, pp. 66-67). These representations are commonly interpreted as an agent's hypotheses, for instance, about hidden causes or the source of a sensory signal. In Bayesian epistemology, the hypotheses are modelled as propositions where their assigned probabilities are interpreted as degrees of beliefs (cf. Talbott, 2015). The probabilities associated with these representations in a Bayesian model are called the 'prior probability' and the 'likelihood'. The prior probability is the agent's representation of how uncertain an event is to occur prior to observing it. The likelihood represents uncertainty about the relationships between those possible events and sensory evidence that could be received.

Morse code. A Morse code is a sequence of signals that can be translated into an alphabet. It can be used to transmit words and sentences across large distances. Figure A illustrates the international Morse Code alphabet and numerals. Signals are encoded as combinations of dots and dashes. A Morse code contains information about syntactic relations within such a sign system and serves the purpose of conveying telegraphic messages. To produce and interpret messages with the code, one can follow steps 1-5 in the figure's legend.



Object perception. This is the ability to find a stable representation on the basis of unstable sensory patterns (e.g., unstable light reflections on the retinal surface) for the purpose of guiding behaviour adaptively (cf. Shepard, 1981/2017). For example, being able to perceive the shape of a cup and to separate its figure from the background enables the perceiver to grasp the cup and drink from it.

Perception as a part of cognition relates to the process of interpreting the objects of sensation in an information-processing procedure with computations on internal representations. This perspective follows the 'sandwich model' (Hurley, 2002), in which perception and action embrace cognition, which links the two like the filling of a sandwich. The key assumption of this view is that perception is the

input for cognition, which sends a motor-output signal to the body to facilitate the appropriate action. The perceived environment and the active body are only peripheral to understanding the nature of cognition. Recently, there has been a lot of debate about the sandwich model, yet, it is still a common ground in current cognitive computational modelling practice.

Perceptual categorisation. Perceptual categorisation (abbrev. 'PC') is a cognitive process by which a set of known perceptual observations is grouped into a mental representation—a perceptual category—that enables an agent to generalise her behaviour appropriately to novel observations. Following Bundesen (1990, p. 523), examples for perceptual categories are "the class of red elements (a color category), the class of letters of type A (a shape category), and the class of elements located at fixation (a location category)." In PC, the 'perceptual' part refers to the notion of a capability for representation (e.g., representing a red colour shade), whereas the 'categorisation' part refers to a decision (e.g., the decision that the red colour shade should be paid attention to).

PC is different from generalisation behaviour because the multiple different ways to categorise may lead to the same generalisation behaviour.

Rational/Task analysis. Bayesian models are often located at the level of computational explanation in Marr's 1982 hierarchy (cf. Colombo & Seriès, 2012; Danks, 2008; Icard, 2018; Zednik & Jäkel, 2016). At this level, the task that a cognitive (i.e., an information-processing) system has to solve is specified, and a rationale for why it has to solve that task is given. Roughly, Bayesian inference models specify cognition as a task of Bayesian inference. The task is to infer the posterior probability of a hypothesis (e.g. about a hidden cause in the world) in light of some piece of evidence (e.g. a sensory signal) from the probability of observing the evidence given that the hypothesis is true together with some background knowledge which is encoded in the probability of the hypothesis regardless of the evidence (where a probability value in the model indicates the agent's degree of belief in the hypothesis). In Bayesian rational analysis, this task is framed under the additional consideration of what would be the optimal behaviour of a rational agent in a given environment, where optimal behaviour is normally considered to be behaviour that is most adaptive to a given environmental niche (cf. Anderson, 1991a). This is then considered useful for systematising and predicting behaviour because it is then possible to ask how the behaviour would change under changes in the given environment (see also van Rooij et al. (2018) on 'what-if' explanations). The (optimal) solution to the task is usually framed as a generic function, typically following Bayes' Theorem.

Probabilistic dependence. The probability of landing 'heads' on a coin toss given that the coin is biased towards heads is different from the probability of

landing 'heads' if the coin was not biased. Knowing that the coin is biased towards 'heads' should affect how probable one is to land a 'head' on the next toss.

Definition A.0.3 (Probabilistic dependence)**.** *Two events or propositions are probabilistically dependent if knowing that one event occurred—or that one proposition is true—does affect the probability of the other event to occur—or of the other proposition to be true. A and B stand for events or propositions (e.g., landing 'heads' and the coin is biased towards heads),*

$Pr(A, B) \neq Pr(A) \times Pr(B)$ if $Pr(A|B) \neq Pr(A)$.

A and B are probabilistically dependent with respect to a probability function Pr if and only if the joint probability of A and B is unequal to the product of their individual probabilities or if the conditional probability of A given B is unequal to the unconditional probability of A.

Probabilistic independence. The probability of throwing a '4' on a die given that it is Wednesday is the same as the probability of throwing a '4' if it was not Wednesday. Knowing that it is Wednesday should not affect whether one is going to throw a '4' on the die.

Definition A.0.4 (Probabilistic independence)**.** *Two events or propositions are probabilistically independent if knowing that one event occurred or that one proposition is true does not affect the probability of the other event to occur or of the other proposition to be true. A and B stand for events or propositions,*

$Pr(A, B) = Pr(A) \times Pr(B)$ if $Pr(A|B) = Pr(A)$.

In words: A and B are probabilistically independent with respect to a probability function Pr if and only if the joint probability of A and B equals the product of their individual probabilities or if the conditional probability of A given B equals the unconditional probability of A.

Product rule (also called 'chain rule'). Says that the joint probability of two events or propositions, E and H, is the product of the probability that one event/proposition has occurred/is true given that the other event/proposition has occurred/is true and the unconditional probability that the given event has occurred. For example, the probability to roll a '4' on a die twice in a row is the product of the probability to roll a '4' on the second throw, given that there was a '4' on the first throw and the probability of having a '4' on the first throw. Formally: $Pr(H, E) = Pr(E|H) \times Pr(H)$.

Bibliography

Anderson, J. R. (1988/2014). The place of cognitive architectures in a rational analysis. In K. V. Lehn (Ed.), *Architectures for intelligence: The 22nd carnegie mellon symposium on cognition* (pp. 13–36).

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409-429.

Anderson, J. R. (1991a). Is human cognition adaptive? *Behavioral and Brain Sciences*, *14*(3), 471–485.

Anderson, J. R., & Matessa, M. (1990). A rational analysis of categorization. In *Machine learning proceedings 1990: Proceedings of the seventh international conference, austin, texas june 21–23* (pp. 76–84). Elsevier.

Baker, A. (2016). Simplicity. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Winter 2016 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2016/entries/simplicity/.

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, *59*(1), 617-645.

Beck, J. (2019). Perception is analog: The argument from weber's law. *The Journal of Philosophy*, *116*(6), 319–349.

Berry, J. W. (1968). Ecology, perceptual development and the müller-lyer illusion. *British journal of Psychology*, *59*(3), 205–210.

Bloom, P. (2002). *How Children Learn the Meanings of Words*. Cambridge, Massachusetts: MIT Press.

Blumson, B. (2012). Mental maps. *Philosophy and Phenomenological Research*, *85*(2), 413–434.

Borg, I., & Groenen, P. J. (2005). *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media.

Boring, E. G. (1943). Sensation and perception in the history of experimental psychology. *Philosophy and Phenomenological Research*, *4*(1), 104-106.

Brighton, H., & Gigerenzer, G. (2008). Bayesian brains and cognitive mechanisms: Harmony or dissonance. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind: Prospects for bayesian cognitive science* (pp. 189–208). New York: Oxford University Press.

Brössel, P. (2015). Keynes's coefficient of dependence revisited. *Erkenntnis*, *80*(3), 521–553.

Bundesen, C. (1990). A theory of visual attention. *Psychological review*, *97*(4), 523.

Camp, E. (2007). Thinking with maps. *Philosophical Perspectives*, *21*(1), 145–182.

Carey, S. (2009). *The origin of concepts*. Ney York: Oxford University Press.

Carnap, R. (1980). A basic system of inductive logic, part ii. *Studies in inductive logic and probability*, *2*, 7–155.

Carnap, R. (1988). *Meaning and necessity: a study in semantics and modal logic*. University of Chicago Press.

Casey, G., & Moran, A. (1989). The computational metaphor and cognitive psychology. *The Irish Journal of Psychology*, *10*(2), 143–161.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive psychology*, *4*(1), 55–81.

Chater, N., & Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends in cognitive sciences*, *3*(2), 57–65.

Chater, N., & Oaksford, M. (2008). *The probabilistic mind: Prospects for a bayesian cognitive science*. New York: Oxford University Press.

Chater, N., & Vitányi, P. M. (2003). The generalized universal law of generalization. *Journal of Mathematical Psychology*, *47*(3), 346–369.

Cheng, P. W., & Pachella, R. G. (1984). A psychophysical approach to dimensional separability. *Cognitive Psychology*, *16*(3), 279–304.

Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use*. Westport: Greenwood Publishing Group.

Clark, A. (2013). Whatever next? predictive brains, situated agents, and the future of cognitive science. *Bahvioral and Brain Sciences*, *36*(3), 1-73.

Colombo, M., & Hartmann, S. (2017). Bayesian cognitive science, unification, and explanation. *The British Journal for the Philosophy of Science*, *68*(2), 451–484. doi: https://doi.org/10.1093/bjps/axv036

Colombo, M., & Seriès, P. (2012). Bayes in the brain—on bayesian modelling in neuroscience. *The British Journal for the Philosophy of Science*, *63*(3), 697–723.

Cooper, L. A., & Shepard, R. N. (1973). Chronometric studies of the rotation of mental images. In *Visual information processing* (pp. 75–176). Elsevier.

Craver, C. F., & Kaplan, D. M. (2018). Are more details better? on the norms of completeness for mechanistic explanations. *The British Journal for the Philosophy of Science*.

Danks, D. (2008). Rational analyses, instrumentalism, and implementations. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind: Prospects for rational models of cognition* (chap. 3). New York: Oxford University Press.

Danks, D. (2014). *Unifying the mind: Cognitive representations as graphical models*. Cambridge, Massachusetts: MIT Press.

Dautriche, I., Chemla, E., & Christophe, A. (2016). Word learning: Homophony and the distribution of learning exemplars. *Language Learning and Development*, *12*(3), 231-251. Retrieved from `http://dx.doi.org/10.1080/15475441.2015.1127163` doi: 10.1080/15475441.2015.1127163

Decock, L., & Douven, I. (2011). Similarity after goodman. *Review of Philosophy and Psychology*, *2*, 61-75.

Decock, L., Douven, I., & Sznajder, M. (2016). A geometric principle of indifference. *Journal of Applied Logic*, *19*, 54–70.

Dennett, D. C. (1987). *The intentional stance*. Cambridge, MA: MIT Press.

Douven, I., & Gärdenfors, P. (2018). What are natural concepts? a design

perspective. *Mind & Language*.

Douven, I., Wenmackers, S., Jraissati, Y., & Decock, L. (2016). Measuring graded membership: The case of color. *Cognitive science*.

Dretske, F. I. (1988). *Explaining behavior: Reasons in a world of causes*. Cambridge, Massachusetts: MIT Press.

Eberhardt, F., & Danks, D. (2011). Confirmation in the cognitive sciences: The problematic case of bayesian models. *Minds and Machines*, *21*(3), 389–410.

Falk, R., & Bar-Hillel, M. (1983). Probabilistic dependence between events. *The Two-Year College Mathematics Journal*, *14*(3), 240–247.

Fodor, J. A. (1975). *The language of thought*. New York: Thomas Y. Crowell.

Fodor, J. A. (1987). *Psychosemantics: The problem of meaning in the philosophy of mind*. Oxford, UK: The MIT Press.

Fodor, J. A. (1998). *Concepts: Where cognitive science went wrong*. Oxford, UK: Claredon Press.

Fodor, J. A. (2008). *Lot2: The language of thought revisited*. New York: Oxford University Press.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1-2), 3–71.

Frank, M., Goodman, N., Lai, P., & Tenenbaum, J. (2009). Informative communication in word production and word learning. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 31).

Frank, S. A. (2018). Measurement invariance explains the universal law of generalization for psychological perception. *Proceedings of the National Academy of Sciences*, *115*(39), 9803–9806.

Gärdenfors, P. (2000). *Conceptual Spaces: The Geometry of Thought*. Cambridge, Massachusetts: MIT Press.

Gärdenfors, P. (2014). *The geometry of meaning: Semantics based on conceptual spaces*. Cambridge, Massachusetts: MIT Press.

Gelman, S. A., & Markman, E. M. (1987). Young children's inductions from natural kinds: The role of categories and appearances. *Child Development*, *58*(6), 1532-1541.

Ghirlanda, S., & Enquist, M. (2003). A century of generalization. *Animal Behaviour*, *66*(1), 15–36.

Goodman, N. (1955). *Fact, fiction and forecast*. Indianapolis: The Bobbs-Merrill Company.

Goodman, N. (1968). *Languages of art: An approach to a theory of symbols*. New York: The Bobbs-Merrill Company.

Goodman, N. (1972). Seven strictures on similarity. In *Problems and projects* (1st print. ed.). Indianapolis: Bobbs-Merrill.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in cognitive sciences*, *14*(8), 357–364.

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition. In R. Sun (Ed.), *The cambridge handbook of computational cognitive modeling*. Cambridge University Press.

Guttman, N., & Kalish, H. I. (1956a). Discriminability and stimulus generaliza-

tion. *Journal of experimental psychology*, *51*(1), 79-88.
Guttman, N., & Kalish, H. I. (1956b). Discriminability and stimulus generalization. *Journal of experimental psychology*, *51*(1), 79.
Harnard, S. (1987). Psychophysical and cognitive aspects of categorical perception: A critical overview. In *Categorical perception: The groundwork of cognition* (chap. 1). New York: Cambridge Univserity Press.
Hartmann, S., & Sprenger, J. (2010). Bayesian epistemology. In S. Bernecker & D. Pritchard (Eds.), *The routledge companion to epistemology* (p. 609-620). London: Routledge.
Haugeland, J. (1981). Analog and analog. *Philosophical Topics*, *12*(1), 213–225.
Hempel, C. G. (1942). The function of general laws in history. *The Journal of Philosophy*, *39*(2), 35–48.
Huber, F. (2016). Formal representations of belief. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Spring 2016 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2016/entries/formal-belief/.
Hurley, S. L. (2002). *Consciousness in action*. Harvard University Press.
Icard, T. F. (2018). Bayes, bounds, and rational analysis. *Philosophy of Science*, *85*(1), 79–101.
Isaac, A. (2018). Computational thought from descartes to lovelace. In M. Sprevak & M. Colombo (Eds.), *The routledge handbook of the computational mind* (pp. 9–22). New York: Routledge.
Isaac, A. M. (2013). Objective similarity and mental representation. *Australasian Journal of Philosophy*, *91*(4), 683–704.
Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*(87-102).
Kitcher, P. (1981). Explanatory unification. *Philosophy of science*, *48*(4), 507–531.
Kitcher, P. (1989). Explanatory unification and the causal structure of the world. In P. Kitcher & W. Salmon (Eds.), *Scientific explanation.* Minneapolis: University of Minnesota Press.
Knill, D. C., & Richards, W. (1996). *Perception as bayesian inference*. New York: Cambridge University Press.
Kosslyn, S. M. (1980). *Image and mind*. Cambridge, MA and London, England: Harvard University Press.
Kouider, S., & Faivre, N. (2017). Conscious and unconscious perception. *The Blackwell Companion to Consciousness*, 551–561.
Krumhansl, C. L. (1978). Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density. *Psychological Review*, *85*(5), 445–463.
Larson, J., Drew, K. L., Folkow, L. P., Milton, S. L., & Park, T. J. (2014). No oxygen? no problem! intrinsic brain tolerance to hypoxia in vertebrates. *Journal of Experimental Biology*, *217*(7), 1024–1039.
Laurence, S., & Margolis, E. (2002). Radical concept nativism. *Cognition*, *86*(1), 25–55.

Leach, E. R. (1964). Anthropological aspects of language: animal categories and verbal insults. In E. H. Lenneberg (Ed.), *New directions in the study of language*. Cambridge, Massachusetts: MIT press.

LeDoux, J. (1998). *The emotional brain: The mysterious underpinnings of emotional life*. New York: Simon and Schuster.

Lewis, D. (1973). *Counterfactuals and comparative possibility*. Springer.

Locke, J. (1805). *An essay concerning human understanding* (The twenty-first edition ed.). London: Clerkenwell Press.

Luce, R., Bush, R. R., & Galanter, E. E. (1963). *Handbook of mathematical psychology: I.* New York: John Wiley.

Machery, E. (2013). In defense of reverse inference. *The British Journal for the Philosophy of Science*, *65*(2), 251–267.

Maddox, W. T. (1992). *Perceptual and decisional separability*. Lawrence Erlbaum Associates, Inc.

Maley, C. J. (2011). Analog and digital, continuous and discrete. *Philosophical Studies*, *155*(1), 117–131.

Margolis, E., & Laurence, S. (1999). *Concepts: Core readings*. MIT Press.

Margolis, E., & Laurence, S. (2011). Learning matters: The role of learning in concept acquisition. *Mind & Language*, *26*(4), 507-539.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York: Freeman.

McGuire, W. J. (1961). A multiprocess model for paired-associate learning. *Journal of Experimental Psychology*, *62*(4), 335.

Melara, R. D. (1992). The concept of perceptual similarity: From psychophysics to cognitive psychology. In D. Algom (Ed.), *Psychophysical approaches to cognition* (Vol. 92, pp. 303–388). Amsterdam: North-Holland.

Miłkowski, M. (2016). Unification strategies in cognitive science. *Studies in Logic, Grammar and Rhetoric*, *48*(1), 13–33.

Millikan, R. G. (1989). Biosemantics. *Journal of Philosophy*, *86*, 281-297.

Müller-Trede, J., Sher, S., & McKenzie, C. R. (2015). Transitivity in context: A rational analysis of intransitive choice and context-sensitive preference. *Decision*, *2*(4), 280.

Murdock, B. (1962). The serial position effect of free recall. *Journal of experimental psychology*, *64*(5), 482.

Myrvold, W. C. (2003). A bayesian account of the virtue of unification. *Philosophy of Science*, *70*(2), 399–423.

Navarro, D. J., Dry, M. J., & Lee, M. D. (2012). Sampling assumptions in inductive generalization. *Cognitive Science*, *36*(2), 187–223.

Newell, A. (1994). *Unified theories of cognition*. Harvard University Press.

Nolan, D. (1997). Quantitative parsimony. *The British Journal for the Philosophy of Science*, *48*(3), 329–343.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of experimental psychology*, *115*(1), 39-61.

Nosofsky, R. M. (1991). Stimulus bias, asymmetric similarity, and classification. *Cognitive Psychology*, *23*(1), 94–140.

Palmeri, T. (2001). The time course of perceptual categorization. In U. Hahn & M. Ramscar (Eds.), *Similarity and categorization* (pp. 193–224). New York: Oxford University Press.

Perfors, A., Tenenbaum, J. B., Griffiths, T. L., & Xu, F. (2011). A tutorial introduction to bayesian models of cognitive development. *Cognition*, *120*(3), 302–321.

Piaget, J. (1964). Cognitive development in children. *Journal of Research in Science Teaching*, *2*(2), 176–186.

Piaget, J. (1976). Piaget's theory. In *Piaget and his school* (pp. 11–23). Springer.

Pinker, S. (2000). Survival of the clearest. *Nature*, *404*(6777), 441–442.

Pinker, S. (2003). Language as an adaptation to the cognitive niche. In *Language evolution* (pp. 16–37). Oxford: Oxford University Press.

Poth, N. L. (2019). Conceptual spaces, generalisation probabilities and perceptual categorisation. In M. Kaipainen (Ed.), *Conceptual spaces: Elaborations and applications* (p. 7-28). Springer.

Poth, N. L., & Broessel, P. (2020). Learning concepts: A learning-theoretic solution to the complex-first paradox. *Philosophy of Science*, *87*(1), 135–151.

Quine, W. V. O. (1948). On what there is. *The review of metaphysics*, *2*(1), 21–38.

Quine, W. V. O. (1960). *Word & Object*. Cambridge, Massachusetts: The MIT Press.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive psychology*, *3*(3), 382–407.

Rosch, E., Mervis, C., Gray, W., Johnson, D., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *7*, 573-605.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive psychology*, *7*(4), 573-605.

Rosch, E. H. (1973). Natural categories. *Cognitive psychology*, *4*(3), 328–350.

Rothkopf, E. Z. (1957). A measure of stimulus similarity and errors in some paired-associate learning tasks. *Journal of Experimental Psychology*, *53*(2), 94.

Sattath, S., & Tversky, A. (1987). On the relation between common and distinctive feature models. *Psychological Review*, *94*(1), 16.

Shanon, B. (1988). On the similarity of features. *New ideas in psychology*, *6*(3), 307–321.

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, *22*(4), 325–345.

Shepard, R. N. (1962). The analysis of proximities: multidimensional scaling with an unknown distance function. i. *Psychometrika*, *27*(2), 125–140.

Shepard, R. N. (1963). Analysis of proximities as a technique for the study of information processing in man. *Human factors*, *5*(1), 33–48.

Shepard, R. N. (1975). Form, formation, and transformation of internal representations. In *Information processing and cognition: The loyola symposium* (pp. 87–122).

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science*, *210*(4468), 390–398.

Shepard, R. N. (1981/2017). Psychophysical complementarity. In M. Kubovy & J. R. Romerantz (Eds.), *Perceptual organization* (pp. 279–341). Oxon/NewYork: Routledge.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317-1323.

Shepard, R. N. (1994). Perceptual-cognitive universals as reflections of the world. *Psychonomic Bulletin & Review*, *1*(1), 2–28.

Shepard, R. N. (2001). Perceptual-cognitive universals as reflections of the world. *Behavioral and brain sciences*, *24*(4), 581–601.

Shepard, R. N., & Arabie, P. (1979). Additive clustering: Representation of similarities as combinations of discrete overlapping properties. *Psychological Review*, *86*(2), 87.

Shepard, R. N., & Chipman, S. (1970). Second-order isomorphism of internal representations: Shapes of states. *Cognitive psychology*, *1*(1), 1–17.

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*(3972), 701–703.

Shu, Y., Hasenstaub, A., Duque, A., Yu, Y., & McCormick, D. A. (2006). Modulation of intracortical synaptic potentials by presynaptic somatic membrane potential. *Nature*, *441*(7094), 761.

Sivik, L., & Taft, C. (1994). Color naming: A mapping in the imcs of common color terms. *Scandinavian journal of psychology*, *35*(2), 144–164.

Sloutsky, V. M. (2010). From perceptual categories to concepts: What develops? *Cognitive science*, *34*(7), 1244–1286.

Smith, L. B., & Samuelson, L. (2006). An attentional learning account of the shape bias: Reply to Cimpian and Markman (2005) and Booth, Waxman, and Huang (2005). *Developmental Psychology*, *42*(6), 1339–1343. doi: https://doi.org/10.1037/0012-1649.42.6.1339

Sperling, G. (1960). The information available in brief visual presentations. *Psychological monographs: General and applied*, *74*(11), 1.

Sprenger, J., & Hartmann, S. (2019). *Bayesian philosophy of science*. Oxford University Press.

Sternberg, R. J. (1980). Representation and process in linear syllogistic reasoning. *Journal of Experimental Psychology: General*, *109*(2), 119.

Stinson, C. (2018). Explanation and connectionist models. In M. Sprevak & M. Colombo (Eds.), *The routledge handbook of the computational mind* (pp. 120–133). New York: Routledge.

Stöckle-Schobel, R. (2012). Perceptual learning and feature-based approaches to concepts – a critical discussion. *Frontiers in Psychology*, *3*(93).

Strevens, M. (2012). Notes on Bayesian Confirmation Theory. In A. Bird & J. Ladyman (Eds.), *Arguing about science* (pp. 293–328). Routledge.

Sun, R. (2008). Introduction to computational cognitive modeling. *Cambridge handbook of computational psychology*, 3–19.

Talbott, W. (2015). Bayesian epistemology. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Summer 2015 ed.).

http://plato.stanford.edu/archives/sum2015/entries/epistemology-bayesian/.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and bayesian inference. *Behavioral and brain sciences*, *24*(4), 629–640.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279-1285.

Turing, A. M. (1936). On computable numbers, with an application to the entscheidungsproblem. *Journal of Masthematics*, *58*(5), 345-363.

Tversky, A. (1977). Features of similarity. *Psychological review*, *84*(4), 327–352.

Tversky, A. (2004). *Preference, belief, and similarity* (E. Shafir, Ed.). MIT Press.

Tversky, A., & Gati, I. (1978). Studies of similarity. *Cognition and categorization*, *1*, 79–98.

van Rooij, I., Wright, C. D., Kwisthout, J., & Wareham, T. (2018). Rational analysis, intractability, and the prospects of 'as if'-explanations. *Synthese*, *195*(2), 491–510.

Volterra, V. (1926). *Fluctuations in the abundance of a species considered mathematically.* Nature Publishing Group.

Von Eckardt, B. (2012). The representational theory of mind. In K. Frankish & W. Ramsey (Eds.), *The cambridge handbook of cognitive science* (pp. 29–50). Cambridge, UK: Cambridge University Press.

Weisberg, M. (2012). *Simulation and similarity: Using models to understand the world*. Oxford University Press.

Wertheimer, M. (1923/2013). Laws of organization in perceptual forms. In W. D. Ellis (Ed.), *A source book of gestalt psychology* (p. 71–88). London: Routledge & Kegan Paul.

Wiese, W. (2017). What are the contents of representations in predictive processing? *Phenomenology and the Cognitive Sciences*, *16*(4), 715–736. doi: 10.1007/s11097-016-9472-0

Wittgenstein, L. (2006). *Philosophische untersuchungen, philosophical investigations* (5th ed.; G. E. M. Anscombe, Trans.). Oxford: Blackwell.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as bayesian inference. *Psychological review*, *114*(2), 245.

Xu, F., & Tenenbaum, J. B. (2007a). Word learning as bayesian inference. *Psychological Review*, *114*(2), 245-272.

Yearsley, J. M., Barque-Duran, A., Scerrati, E., Hampton, J. A., & Pothos, E. M. (2017). The triangle inequality constraint in similarity judgments. *Progress in biophysics and molecular biology*, *130*, 26–32.

Zednik, C., & Jäkel, F. (2014). How does bayesian reverse-engineering work? In *Proceedings of the cognitive science society* (Vol. 36).

Zednik, C., & Jäkel, F. (2016). Bayesian reverse-engineering considered as a research strategy for cognitive science. *Synthese*, *193*(12), 3951–3985.